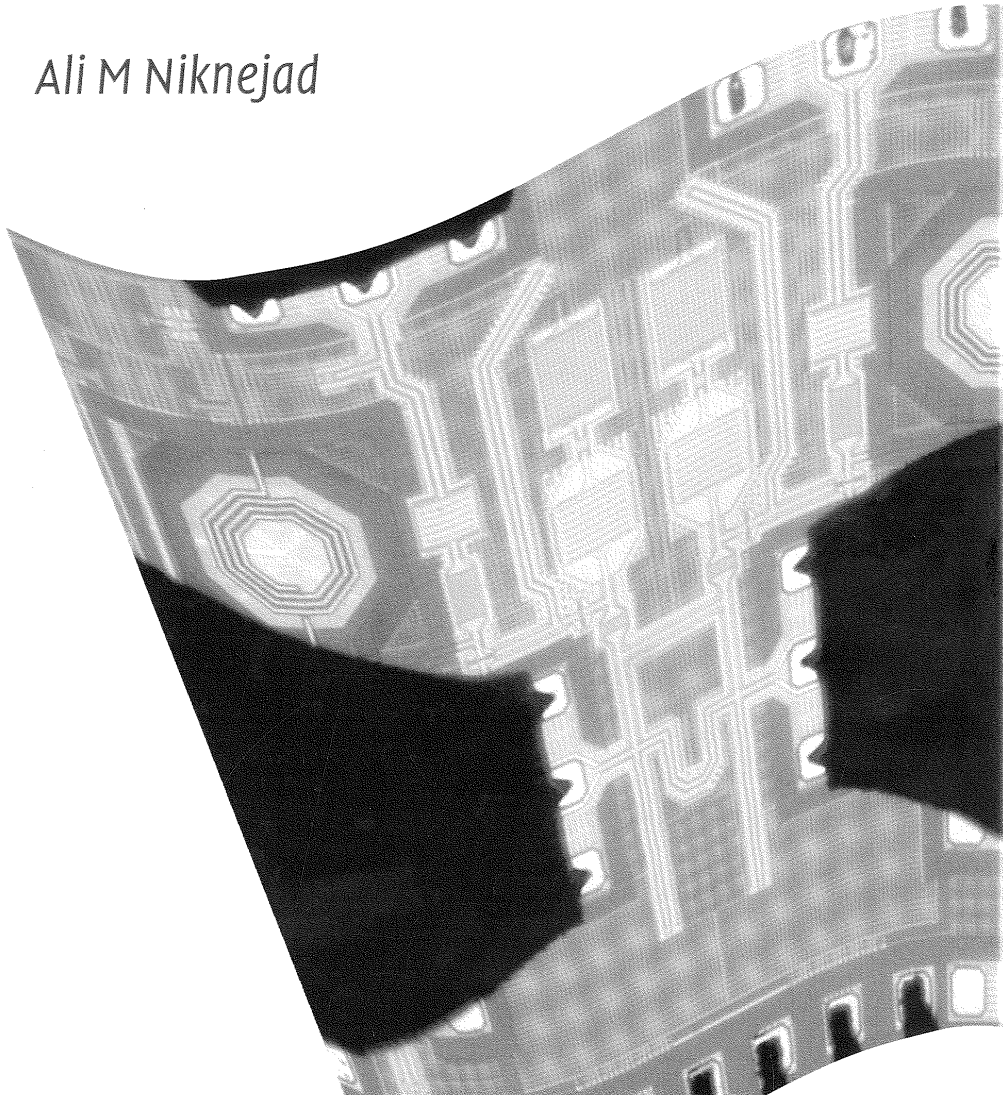


Electromagnetics

for High-Speed Analog *and*
Digital Communication Circuits

Ali M Niknejad



Niknejad

Electromagnetics for High-Speed Analog
and Digital Communication Circuits

CAMBRIDGE

Electromagnetics for High-Speed Analog and Digital Communication Circuits

Modern communications technology demands smaller, faster, and more efficient circuits, the design of which requires a good understanding of circuit theory and electromagnetics. This book reviews the fundamentals of electromagnetism as applied to passive and active circuit elements, highlighting the various effects and potential problems in designing a new circuit. The author begins with a review of the basics: the origin of resistance, capacitance, and inductance, from a circuit and field perspective; then progresses to more advanced topics such as passive device design and layout, resonant circuits, impedance matching, high-speed switching circuits, and parasitic coupling and isolation techniques. Using examples and applications in RF and microwave systems, the author describes transmission lines, transformers, and distributed circuits. State-of-the-art developments in Si-based broadband analog, RF, microwave, and mm-wave circuits are also covered. With up-to-date results, techniques, practical examples, many illustrations, and worked examples, this book will be valuable to advanced undergraduate and graduate students of electrical engineering and practitioners in the IC design industry. Further resources for this title are available at www.cambridge.org/9780521853507.

ALI M. NIKNEJAD obtained his Ph.D. in 2000 from the University of California, Berkeley, where he is currently an associate professor in the EECS department. He is a faculty director at the Berkeley Wireless Research Center (BWRC) and the co-director of the BSIM Research Group. Before his appointment at Berkeley, Niknejad worked for several years in industry designing CMOS and SiGe ICs. He has also served as an associate editor of the *IEEE Journal of Solid-State Circuits*, and was a co-recipient of the Jack Raper Award for Outstanding Technology Directions Paper at ISSCC 2004.

Electromagnetics for High-Speed Analog and Digital Communication Circuits

ALI M. NIKNEJAD



CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi - 110025, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9780521853507

© Cambridge University Press 2007

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2007

Reprinted with corrections 2008

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-85350-7 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

<i>Preface</i>	<i>page ix</i>
<i>Acknowledgments</i>	<i>xi</i>
1 Introduction	1
1.1 Motivation	1
1.2 System in Package (SiP): chip and package co-design	13
1.3 Future wireless communication systems	13
1.4 Circuits and electromagnetic simulation	15
2 Capacitance	18
2.1 Electrostatics review	18
2.2 Capacitance	32
2.3 Non-linear capacitance	41
2.4 References	52
3 Resistance	53
3.1 Ohm's Law	53
3.2 Conduction in semiconductors	59
3.3 Diffusion	66
3.4 Thermal noise	68
3.5 References	73
4 Ampère, Faraday, and Maxwell	74
4.1 Ampère: static magnetic fields	74
4.2 Magnetic materials	82
4.3 Faraday's big discovery	88
4.4 Maxwell's displacement current	91
4.5 References	95
5 Inductance	96
5.1 Introduction	96
5.2 Inductance	97
5.3 Magnetic energy and inductance	101
5.4 Discussion of inductance	106
	v

5.5	Partial inductance and return currents	119
5.6	Impedance and quality factor	120
5.7	Frequency response of inductors	121
5.8	Quality factor of inductors	130
5.9	Inductors and switching circuits	133
5.10	Preview: how inductors mutate into capacitors	135
5.11	References	136
6	Passive device design and layout	137
6.1	Ring inductor	137
6.2	The classic coil	141
6.3	Spirals	143
6.4	Symmetric inductors	145
6.5	Multilayer inductors	147
6.6	Inductor equivalent circuit models	149
6.7	Integrated capacitors	150
6.8	Calculation by means of the vector potential	153
6.9	References	165
6.10	Appendix: Filamental partial mutual inductance	165
7	Resonance and impedance matching	168
7.1	Resonance	168
7.2	The many faces of Q	180
7.3	Impedance matching	186
7.4	Distributed matching networks	199
7.5	Filters	199
7.6	References	200
8	Small-signal high-speed amplifiers	201
8.1	Broadband amplifiers	202
8.2	Classical two-port amplifier design	220
8.3	Transistor figures of merit	242
8.4	References	244
9	Transmission lines	246
9.1	Distributed properties of a cable	246
9.2	An infinite ladder network	248
9.3	Transmission lines as distributed ladder networks	249
9.4	Transmission line termination	253
9.5	Lossless transmission lines	255
9.6	Lossy transmission lines	260
9.7	Field theory of transmission lines	264
9.8	T-line structures	265
9.9	Transmission line circuits	272

9.10	The Smith Chart	282
9.11	Transmission line-matching networks	287
9.12	References	292
10	Transformers	293
10.1	Ideal transformers	293
10.2	Dot convention	294
10.3	Coupled inductors as transformers	295
10.4	Coupled inductor equivalent circuits	296
10.5	Transformer design and layout	299
10.6	Baluns	301
10.7	Hybrid transformer	302
10.8	Transformer parasitics	305
10.9	Transformer figures of merit	305
10.10	Circuits with transformers	310
10.11	References	319
11	Distributed circuits	320
11.1	Distributed RC circuits	320
11.2	Transmission line transformers	325
11.3	FETs at high frequency	332
11.4	Distributed amplifier	335
11.5	References	342
12	High-speed switching circuits	343
12.1	Transmission lines and high-speed switching circuits	343
12.2	Transients on transmission lines	345
12.3	Step function excitation of an infinite line	346
12.4	Terminated transmission line	348
12.5	Reactive terminations	357
12.6	Transmission line dispersion	360
12.7	References	363
13	Magnetic and electrical coupling and isolation	364
13.1	Electrical coupling	364
13.2	Magnetic coupling	367
13.3	Ground noise coupling	373
13.4	Substrate coupling	378
13.5	Package coupling	383
13.6	References	385
14	Electromagnetic propagation and radiation	386
14.1	Maxwell's equations in source-free regions	386
14.2	Penetration of waves into conductors	390

14.3	Poynting vector	395
14.4	EM power carried by a plane wave	397
14.5	Complex Poynting Theorem	399
14.6	Reflections from a perfect conductor	402
14.7	Normal incidence on a dielectric	404
14.8	References	406
15	Microwave circuits	407
15.1	What are microwave circuits?	407
15.2	Microwave networks	409
15.3	Lorentz reciprocity theorem	409
15.4	The network formulation	412
15.5	Scattering matrix	414
15.6	Properties of three-ports	421
15.7	Properties of four-ports	429
15.8	Two conductor coupler	438
15.9	References	440
	<i>References</i>	441
	<i>Index</i>	445

Preface

Why another EM book? There are virtually thousands of books written on this subject and yet I felt the urge to write another one.

The idea for this book germinated in my mind on a long and uneventful drive from Berkeley to San Diego. I had just completed my first year of graduate school at Berkeley and had started a research project on analyzing spiral inductors. It occurred to me that studying electromagnetics as a circuit designer was a lot easier than studying it as an undergraduate at UCLA. Even though I took many EM courses during my undergraduate education, very little of it actually stuck with me. Much like all those foreign languages we learn in high school or college, without any practice, we quickly lose our skills. When we find ourselves at that critical moment in a foreign country, our language skills fail us. While EM is the foundation of much of electrical engineering, somehow it's treated as a foreign tongue, spoken only by the few learned folks in the the field. But learning EM should not be like learning Greek or Latin!

That summer I spent many weekends in San Diego visiting my family. During these trips I'd take my EM books down to the beach and study. I'd plant myself on the beach at La Jolla or Del Mar and work my way through my undergraduate EM text. This time around, things were making a lot more sense, since I had an urgent need to actually learn electromagnetics. But I observed that having a circuits background was somewhat equivalent to speaking a related derived tongue. I realized that many people out there also missed the boat on learning EM, since they learned it without any background, desire, or need to learn it. But many of those same people, after taking a lot of high-frequency electronics courses, feel they need to relearn this important subject. If you're one of those people, this book is written for you!

When I was an undergraduate student, EM courses were a required part of every EE student's education. No matter how painful, you had to work your way through two or three courses. But today the situation has changed dramatically. Many schools have made this an optional course and, much to our horror, many students simply skip it! Even though they do take EM as part of their physics education, the emphasis is on fundamentals, with no coverage of important engineering topics such as transmission lines or waveguides. Today, more than ever, this seems like a tragedy. High-speed digital, RF, and microwave circuits abound, necessitating the training of engineers in the art and science of electronics, electromagnetics, communication circuits, antennas, propagation, etc.

With the availability of high-speed 64-bit microprocessors, server farms, Gb/s networks, and mass storage, many practical problems are now computationally tractable. Workers in the field of high-speed electronics are increasingly turning to commercial electromagnetic

solvers to tackle difficult problems. As powerful as EM solvers are today, it still takes a lot of skill to set up and run a problem. And at the end of a long five hour simulation, can you trust the results? Did you actually set up the problem correctly? Are the boundary conditions appropriate? Is the field accuracy high enough? These are difficult questions and can only be answered by observing the currents, voltages, and electric and magnetic fields with a trained eye.

The focus of this book is the application of electromagnetics to circuit design. In contrast to classical analog integrated circuit design, passive components play an integral role in the design of RF, microwave, and broadband systems. Most books dedicate a section or at best a chapter to this all important topic.

The book begins with the fundamentals – the origins of resistance, capacitance, and inductance. We spend a great deal of time reviewing these fundamental passive elements from a circuit and field perspective. With this solid foundation, the book progresses to more advanced applications. A chapter on passive device design and layout reviews state-of-the-art layout techniques for the realization of passive devices in an integrated circuit environment. Important circuit applications such as resonant circuits and impedance matching are covered extensively with an emphasis on the inner workings of the circuitry (rather than a cookbook approach) in order to uncover important insights into the insertion loss of these circuits. Next, the book moves to active two-port circuits and reviews the co-design of amplifiers with passive components. Two-port circuit theory is used extensively to understand optimal power gain, stability, activity, and unilateral gain. Transmission lines, transformers, and distributed circuits form the core of the advanced circuit applications of passive elements. These topics are taught in a coherent fashion with many important examples and applications to RF and microwave systems. The time-domain perspective is covered in a chapter on high-speed switching circuits, with a detailed discussion of the transient waveforms on transmission lines and transmission line dispersion. Parasitic coupling and isolation techniques are the topic of an entire chapter, including discussion of package, board, and substrate coupling. An introduction to the analysis and design of passive microwave circuits is also covered, serving as a bridge to an advanced microwave textbook.

Acknowledgments

I would like to thank all the people who have helped me write this book. Much of this material was inspired by teaching courses at Berkeley and so I thank all the students who read the original lecture notes and provided feedback in EECS 105, 117, 142, 217, and 242 (thanks to Ke Lu for detailed feedback). This book would not be as interesting (assuming you find it so) without real circuit applications drawn from literature and from our own research projects. Thanks to my colleagues and collaborators at Berkeley who have created a rich and stimulating research environment. In particular, thanks to my BWRC colleagues, Robert Brodersen, Jan Rabaey, Bora Nikolic, Robert Meyer, Paul Wright, and John Wawrzyniek. And thanks to Professor Chenming Hu for inviting me to be a part of the world-famous BSIM team. Thanks to Jane Xi for her hard work and dedication to the BSIM team. Special thanks goes to the graduate student researchers. In particular, thanks to Sohrab Emami and Chinh Doan who were key players in starting the Berkeley 60 GHz project and OGRE. Many of the high-frequency examples come from our experience with this project. Thanks to Professor Andrea Bevilacqua (University of Padova, Italy) for a stimulating research collaboration on UWB. Thanks to Axel Berny and his love of oscillators.

Though I take responsibility for any errors in the book, I have my graduate students to thank for the countless errors they were able to find by reading through early drafts of the manuscript. Thanks to Ehsan Adabi, Bagher (Ali) Afshar, Mounir Bohsali, Yuen Hui Chee, Wei-Hung Chen, Debo Chowdhury, Mohan Dunga, Gang Liu, Peter Haldi, Babak Heydari, and Nuntachai Poobuapheun. They provided detailed feedback on various chapters of the book.

Also thanks to my friends and colleagues for reviewing the book. In particular I'm grateful to Dr. Manolis Terrovitis, Eric Hoffman, Professor Hui Wu, and Professor Hossein Hashemi for taking the time to review the book and provide feedback.

Finally, thanks to the folks who supported our research during the past four years. Special thanks to DARPA and the TEAM project, in particular thanks to Barry Perlman and Dan Radack for your support of university research. Thanks to BWRC member companies, in particular ST Microelectronics, Agilent Technologies, Infineon, Conexant Systems, Cadence, and Qualcomm. Thanks to Analog Devices, Broadcom, Berkeley Design Automation, and National Semiconductor for your support through the UC MICRO and UC Discovery programs. And thanks to SRC and member companies for supporting research of compact modeling at Berkeley. Thanks in particular to Jim Hutchby of SRC, Keith Green of Texas Instruments, Weidong Liu of Synopsys, Judy An of AMD, Josef Watts and Jack Pekarik of IBM, and Ben Gu of Freescale.

1 Introduction

1.1 Motivation

The history of electronics has been inextricably linked with the growth of the communications industry. Electronic communication served as a major enabling technology for the industrial revolution. When scientists and engineers learned to control electricity and magnetism, it did not take long for people to realize that the electromagnetic force would enable long-range communication. Even though the basic science of Maxwell's equations was well understood, it took much longer for practical applications to fully exploit all the fantastic possibilities such as radio, television, and personal wireless communication.

At first only crude wires carrying telegraph signals were rolled out sending Morse code,¹ digital signals at speeds limited by human operators. In this regard it is ironic that digital communication predates analog communication. Telegraph wires were laid alongside train tracks, making long-range communication and transportation a practical reality. Sending signals faster and further ignited the imagination of engineers of the time and forced them to study carefully and understand the electromagnetic force of nature. Today we are again re-learning and inventing new digital and analog communication systems that are once again compelling us to return to the very fundamental science of electricity and magnetism.

The topic of this book is the high-frequency electromagnetic properties of passive and active devices. For the most part, passive devices are resistors, capacitors, transformers, and inductors, while active devices are transistors. Most applications we draw from are high-frequency circuits. For example, radio frequency (RF) circuits and high-speed digital circuits both depend on a firm understanding of passive devices and the environment in which they operate.

Circuit theory developed as an abstraction to electromagnetics. Circuit theory is in effect the limit of electromagnetics for a circuit with negligible dimension. This allows spatial variations and time delay to be ignored in the analysis of the circuit. As such, it allowed practicing engineers to forego solving Maxwell's equations and replaced them with simple concepts such as KCL and KVL. Even differential equations were eliminated and replaced with algebraic equations by employing Laplace transforms. The power and popularity of circuit theory was due to its simplicity and abstraction. It allowed generations of engineers to

¹ Or as Paul Nahin suggests in [41] we should more correctly call this "Vail" code.

solve difficult problems with simple and yet powerful tools. In effect, it allowed generations of engineers to forego reading a book such as this one.

So why read another book on electromagnetics? Why bother learning all this seemingly complicated theory when your ultimate goal is to build circuits and systems for communication and information management?

We live today at the intersection of several interesting technologies and applications. Integrated circuit technology has enabled active devices to operate at increasingly higher frequencies, turning low-cost Si technology into a seemingly universal panacea for a wide array of applications. CMOS digital circuits are switching at increasingly higher rates, pushing multi GHz operation. Si CMOS, bipolar, and SiGe technology have also enabled a new class of low-cost RF and microwave devices, with ubiquitous deployment of cellular phones in the 800 MHz–2 GHz spectrum, and high-speed wireless LAN in the 2–5 GHz bands. There seems to be very little in the way of enabling Si technology to exploit the bandwidths up to the limits of the device technology. In a present-day digital 130 nm CMOS process, for instance, circuits are viable up to 60 GHz [55] [13].

At the same time, wired communication is pushing the limits. Gigabit Ethernet and high-speed USB cables are now an everyday reality, and people are already pursuing a 10 Gb/s solution. Optical communication is of course at the forefront, with data rates in the 40 Gb/s range now commercially viable and at relatively low cost.

The simultaneous improvement in active device technology, miniaturization, and a host of new applications are the driving force of today's engineering. As integrated circuits encompass more functionality, many traditionally off-chip components are pushed into the IC or package, blurring the line between active devices and circuits and passive devices and electromagnetics. This is the topic of this book.

Technology enhancements

The limitations in frequency and thus speed of operation is usually set by the active device technology. One common figure of merit for a technology is, the unity-gain frequency f_T , the frequency at which the short-circuit current gain of the device crosses unity. Another important figure of merit is f_{max} , the maximum frequency of oscillation, or equivalently the frequency where the maximum power gain of a transistor drops to unity. Since f_{max} is a strong function of layout and parasitics in a process, it is less often employed. In contrast, the f_T depends mostly on the dimensions of the transistor and the transconductance

$$f_T = \frac{1}{2\pi} \frac{g_m}{C_\pi + C_\mu} \quad (1.1)$$

It can be shown [50] that the device f_T is inversely related to the transistor dimensions. For a long-channel MOSFET the key scaling parameter is the channel length L

$$f_T \approx \frac{1}{2\pi} \frac{3\mu(V_{gs} - V_t)}{2L^2} \quad (1.2)$$

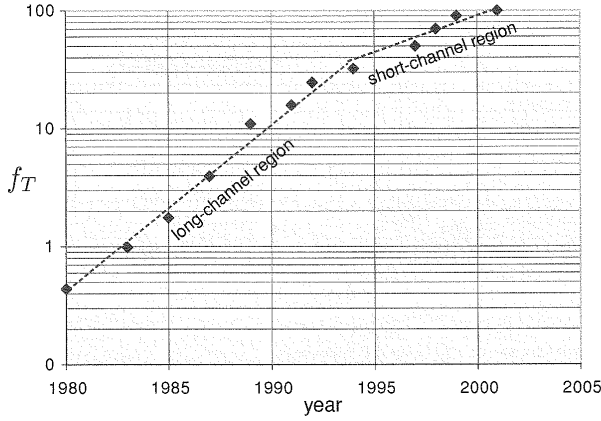


Figure 1.1 The improvements in device unity-gain frequency f_T over the past two decades due to device scaling.

while in the limit for short channel transistors the scaling changes to L^{-1} since the current is limited by velocity saturation

$$I_{ds,sat} = WQ_i v_{sat} = WC_{ox}(V_{gs} - V_t)v_{sat} \quad (1.3)$$

resulting in

$$f_T \propto \frac{v_{sat}}{L} \quad (1.4)$$

For a bipolar junction transistor (BJT) the critical dimension is the base width. In the limit that base transit limits the frequency of operation

$$f_T \approx \frac{1}{2\pi} \frac{1}{\tau_B} \propto \frac{1}{W_B^2} \quad (1.5)$$

As integrated circuit manufacturing technology has improved exponentially in the past three decades, so has the f_T of the device, giving circuit designers increasingly faster devices. A plot of the device f_T over the years for a MOSFET device is shown in Fig. 1.1, and the exponential growth in technological advancements can be seen clearly.

It is important to note that this improvement in performance only applies to the intrinsic device. Early circuits were in fact limited by the intrinsic transistor and not the parasitic routing and off-chip environment. As circuit technology has advanced, though, the situation has reversed and now the limitation is set by the parasitics of the chip and board environment, as well as the performance of the passive devices. This is why the material of this book is now particularly relevant. It can be shown that a good approximation to the CMOS device f_{max} is given by [42]

$$f_{max} \approx \frac{f_T}{2\sqrt{R_g(g_m C_{gd}/C_{gg}) + (R_g + r_{ch} + R_s)g_{ds}}} \quad (1.6)$$

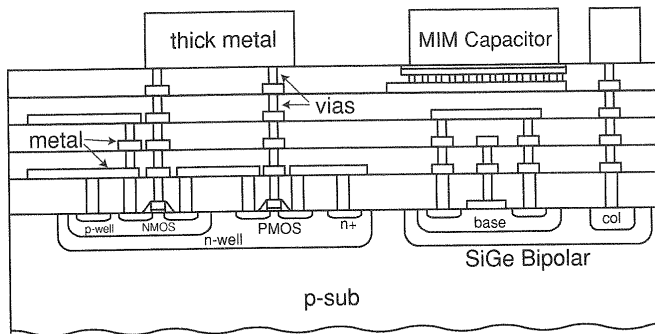


Figure 1.2 Cross section of a SiGe BiCMOS process.

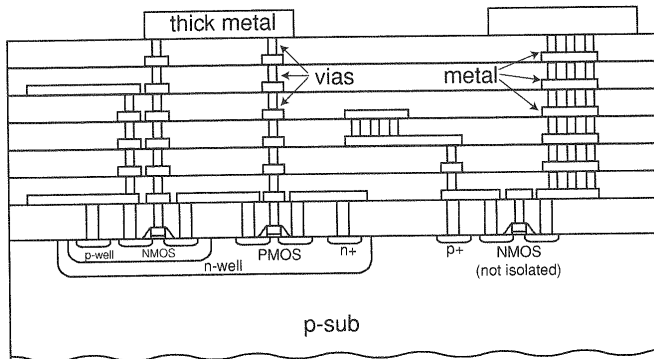


Figure 1.3 Cross section of an advanced CMOS process.

where the device performance is a strong function of the loss, such as the drain/source resistance R_s , R_d , and the gate resistance R_g . These parasitics are in large part determined by layout and the process technology.

While early integrated circuit technologies were limited to a few types of different active devices and a few layers of aluminum interconnect metal, present-day process technology has a rich array of devices and metal routing. In an advanced Si process, shown in Fig. 1.2, high-performance SiGe HBT devices are complemented by MOS and PN-junction varactors, metal-insulator-metal (MIM) high-density and high-quality capacitors, and thick-metal for low-loss interconnect and inductors/transformers. Even a digital CMOS process, as shown in Fig. 1.3, has many advanced capabilities. In addition to several flavors of MOS active devices (fast thin oxide, thick oxide, high/low V_T), there are also enhanced isolation structures and triple-well (deep n-well) devices, and many layers of interconnect that allow construction of high-quality, high-density capacitors and reasonably high-quality inductors.

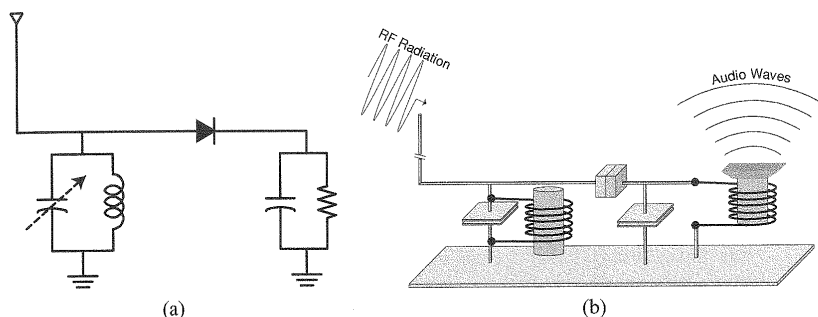


Figure 1.4 (a) A simple AM receiver circuit. The resistor represents a high-input impedance earphone. (b) A physical realization of the simple AM receiver circuit.

Radio and wireless communication

Early radio systems were essentially all passive. To see this look into the back of an old radio where a few active devices (vacuum tubes or transistors) are surrounded by tens to hundreds of passive devices. Consider the circuit diagram of a very simple AM receiver shown in Fig. 1.4a. The antenna drives a resonant tank tuned to the center frequency of the transmitting station. This signal is fed into a peak detector that follows the peak of the RF signal. The low-pass filter time constant is only fast enough to follow the low-frequency audio signal (generically the baseband signal) and yet too slow to follow the RF, thus removing the RF signal and retaining the low-frequency audio. This received signal is usually too weak to drive a speaker but can be heard through a sensitive headphone. A simple audio amplifier can be used to strengthen the signal.

It is interesting to note that this AM receiver can be physically realized by merely using contacts between a few different pieces of metal and semiconductors. This is shown in Fig. 1.4b. The resonant tank is simply a piece of wire wound into a coil which contacts with the capacitor, two metal plates in close proximity. The diode can be realized as the junction of a metal and semiconductor. Finally, to convert electric energy into sound we can use another large inductor coil and use the time-varying magnetic force to move a paper thin cone driven by a magnetic core. Magnetic materials have been known since ancient times and therefore since the metal age we have had the capability to build radio receivers! In fact, it is not surprising that radios often crop up accidentally.²

Most modern radios operate based on an architecture invented by Edwin Armstrong. The block diagram of such a system, called a super-heterodyne receiver, is shown in Fig. 1.7. This receiver incorporates a local oscillator (LO), a block that primarily converts DC power into RF power at the oscillation frequency. A mixer takes the product of this signal and the

² For instance my old answering machine also picked up the radio. Sometimes you could hear it as you were waiting for the tape recorder to rewind. At least this was a desirable parasitic radio.

signal received by the antenna. Recall the following trigonometric identity³

$$2 \cos(\omega_{LO}t) \cos(\omega_{RF}t) = \cos((\omega_{LO} + \omega_{RF})t) + \cos((\omega_{LO} - \omega_{RF})t) \quad (1.7)$$

Note that the product of the received RF signal and the local oscillator signal produces two new signals, one centered at the difference frequency and one centered at the sum frequency. If we put a bandpass filter at one of these frequencies, call it the intermediate frequency, IF, we can electronically tune the radio by simply changing the LO frequency. This is accomplished by using a frequency synthesizer (a PLL or phase locked loop), and thus we avoid building a variable filter common to the early radios. The important point is that the IF is fixed and we can build a very selective filter to pinpoint our desired signal and to reject everything else. Why not simply set LO equal to RF to move everything to DC? This is in fact the direct-conversion or zero-IF architecture. It has some shortcomings such as problems with DC offset,⁴ but its main advantage is that it lowers the complexity of the RF section of a typical radio.

At the heart of the frequency synthesizer is the voltage controlled oscillator (VCO). The VCO is an oscillator where the output frequency is a function of a control voltage or current.⁵ To build a VCO we need a way to change the center frequency of a resonant tank. The resonant tank is simply an inductor in series or in parallel with a capacitor. One typical realization is to use varactor, a variable capacitor. A reversed biased diode serves this purpose nicely, as the depletion region width, and thus the small-signal capacitance, is a function of the reverse bias. It seems that a super-heterodyne receiver has simply moved the variable resonant tank from the antenna front end to a variable resonant tank in the VCO! Have we gained anything? Yes, because the frequency of the VCO can be controlled precisely in a feedback loop (using an accurate frequency reference such as a crystal), eliminating any problems associated with absolute tolerances in components in addition to drift and temperature variation.

The radio has once again emerged as a critical application of passive devices spawned by the growth and popularity of wireless telephones, in particular the cellular phone. By limiting the transmitter powers and taking advantage of spatial diversity (re-using the same frequency band for communication for points far removed – for non-adjacent cell sites), a few hundred radio channels can be used to provide wireless communication to millions of people. Modern cell phones employ complicated radio receivers and transmitters (transceivers) employing hundreds and thousands of passive devices. Early cell phones used simple architectures such as the super-heterodyne receiver but the demand for low-cost and small footprints has prompted a re-investigation of radio architectures.

The layout of a modern 2.4 GHz transceivers for 802.11b wireless LAN (WLAN) is shown in Fig. 1.5 [7]. The IC is implemented in a 0.25μ CMOS process and employs several integrated passive devices such as spiral inductors, capacitors, and resistors. The spiral inductors comprise a large fraction of the chip area. The next chip shown in Fig. 1.6

³ I recall asking my trig teacher about the practical application of the subject. After scratching her head and pondering the question, her response was that architects use trig to estimate the height of buildings! A much better answer would have been this equation.

⁴ And $1/f$ noise in MOS technology.

⁵ This makes a nice AM to FM modulator, as well.

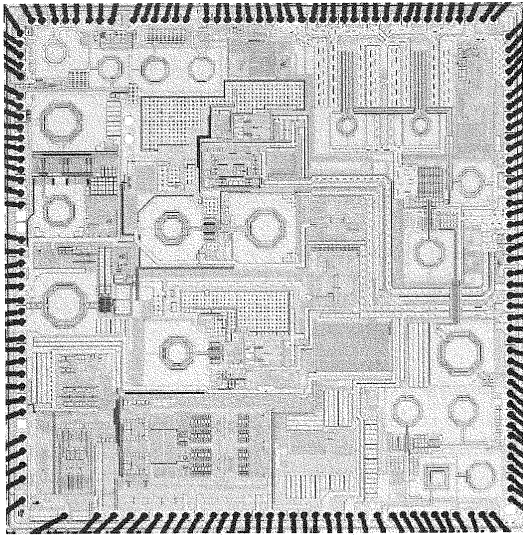


Figure 1.5 A 2.4 GHz CMOS 802.11b Wireless LAN Transceiver [7]. (Copyright 2003, IEEE)

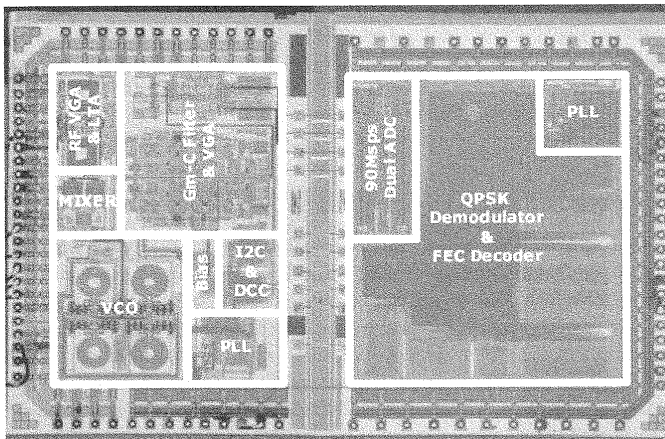


Figure 1.6 A direct-conversion satellite broadband tuner-demodulator SOC [17] operates from 1–2 GHz. (Copyright 2003, IEEE)

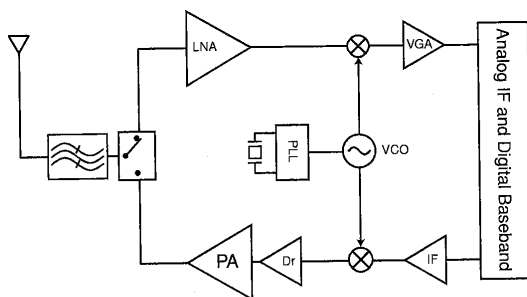


Figure 1.7 The block diagram of an Armstrong super-heterodyne transceiver.

[17] is an integrated direct-conversion satellite broadband tuner-demodulator “system-on-a-chip” (SOC). The chip is implemented in a $0.18\ \mu$ CMOS process and employs MIM capacitors and spiral inductors. It operates in the 1–2 GHz band, requiring broadband operation and high linearity. Notice that the digital baseband has been integrated on to a single chip along with the sensitive analog and RF blocks. This brings about several important challenges in the design due to the parasitic coupling between the various blocks. A triple-well process and lead-less package technology are used to maximize the isolation.

In general, integrating an entire transceiver on to a single chip has many challenges. The power amplifier (PA) or PA driver can injection lock the VCO through the package and substrate, causing a spurious modulation. Digital circuitry can couple seemingly random switching signals into the analog path, effectively increasing the noise floor of the sensitive RF and analog blocks. As the level of integration increases, a single chip or package may contain several systems in operation simultaneously, requiring further understanding and modeling of the coupling mechanisms.

Computers and data communication

Computers and data communication, particularly the Internet, have given rise to a new tidal wave in the information revolution. The speed of computers has improved drastically due to technological improvements in transistor, microprocessor, memory, and system bus architectures. Computer circuits move and process discrete time signals at a frequency determined by the system clock. For instance, in the current generation of computers the clock speed inside the microprocessor is several GHz, while the speed of the system bus and memory lag behind by a factor of 2–3. This is because inside the microprocessor everything is small and dense and signals travel short distances in the presence of small parasitics (mainly capacitance). Off-chip, though, the system bus environment is characterized by much longer distances and much larger parasitics, such as non-ideal dispersive transmission lines along the board traces. Modern computer networks, like gigahertz Ethernet LAN, also

operate at high frequencies over wires, necessitating a complete understanding of distributed transmission line effects. These topics are covered in Chapters 9 and 12.

High-speed wireless data communication is the focus of much research and development. The next and future generations of cellular technology will bring the Internet from our homes and offices into virtually every location on earth. Wireless LAN systems enable short-range high-speed data communication without the expensive network infrastructure. A physical network infrastructure requires time-consuming distribution of cables to every office in a building. A wireless system can be up and running in minutes or hours as opposed to days or months.⁶

In such systems cost and size will force many external passive components on to the chip environment, where knowledge of parasitic coupling and loss is critical in a successful low-cost implementation. In this book we spend a great deal of time discussing inductors, capacitors, transformers, and other key passive elements realized in the on-chip environment.

Microwave systems

Microwave systems employ higher frequencies where the wavelength $\lambda = c/f$ is of the order of centimeters or millimeters. Thus the lumped circuit approach fails since these structures are a significant fraction of a wavelength and spatial variation begins to play as important a role as time variation. Such systems were first employed in World War II for radar systems.⁷ In a radar system, the small wavelength allows us to construct a highly directional antenna to focus a beam of radiation in a given direction. By observing the reflection, we can compute the time-of-flight and hence the distance to an object. By also observing the Doppler frequency shift, we can compute the speed of the object.

Perhaps the greatest difficulty in designing microwave systems below 10 GHz is that the operating frequency is in an intermediate band where lumped element circuit techniques do not strictly apply and microwave methodology results in prohibitively large circuits. At 3 GHz, the wavelength is 10 cm in air and about 5 cm in silicon dioxide, while an integrated circuit has dimensions of the order of millimeters, thus precluding distributed elements such as quarter-wave transmission lines. But using advances pseudo-lumped passive devices such as inductors, transformers, and capacitors, microwave ICs can be realized with minimal off-chip components.

Many early microwave systems were designed for military applications where size and cost were of less concern in comparison to the quality and reliability. This led to many experimental and trial-and-error design approaches. Difficult system specifications were met by using the best available technology, and often expensive and exotic processes were employed to fabricate high-speed transistors. New microwave systems, in contrast, need to be mass produced and cost and size are the main concerns. Fortunately high-volume process

⁶ I seem to recall that it took a year for a network upgrade to occur in Cory Hall at Berkeley!

⁷ It is ironic that the EEs of the time lacked the necessary skills to build such systems and the project was handed off to the physicists at the MIT Radiation Lab.

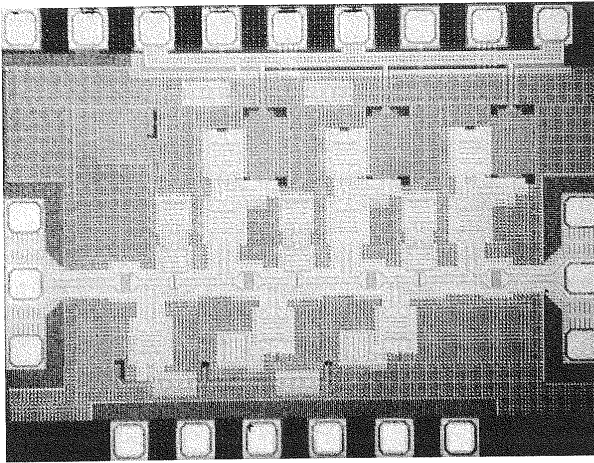


Figure 1.8 A three-stage 60 GHz CMOS LNA implemented in a digital 130 nm process.

technology using silicon is now readily available. The speed is now sufficient to displace many specialized technologies. Since high-volume microwave systems are primarily being designed by circuit engineers as opposed to microwave engineers, the lack of knowledge of electromagnetics and distributed circuits can be an impediment to successful integration and implementation.

Higher-frequency bands offer new opportunities to exploit sparsely used spectrum. The 60 GHz “oxygen absorption” band is a prime example, providing 7 GHz of unlicensed bandwidth in the US. An example of a 60 GHz multi-stage low-noise amplifier (LNA) is shown in Fig. 1.8. Here transmission lines play a key role as inductors, interconnect, and resonators. A 60 GHz single-transistor mixer, shown in Fig. 1.9, employs a hybrid coupler (see Section 15.7) to combine the RF and LO signal. Spiral inductors are also employed in the IF stages. Both of these chips were fabricated in a digital 130 nm CMOS process. Another CMOS microwave circuit is shown in Fig. 1.10. This is a circular standing-wave 10 GHz oscillator, employing integrated transmission lines in the resonator [11]. There is a beautiful connection between this oscillator and the orbit of an electron in a hydrogen atom. Similar to the wave function of an electron, the electromagnetic mode must satisfy the periodic boundary condition, and this determines the possible resonant modes of the structure.

Optical communication

Fiber-optic communication systems allow large amounts of data to be transmitted great distances with relatively little attenuation. At optical frequencies, metals are too lossy for long-haul communication without amplification, and so the energy is confined inside a thin fiber of glass by total internal reflection. The flexibility and low cost of this material has displaced more traditional waveguides made of rigid or semi-rigid and expensive materials.

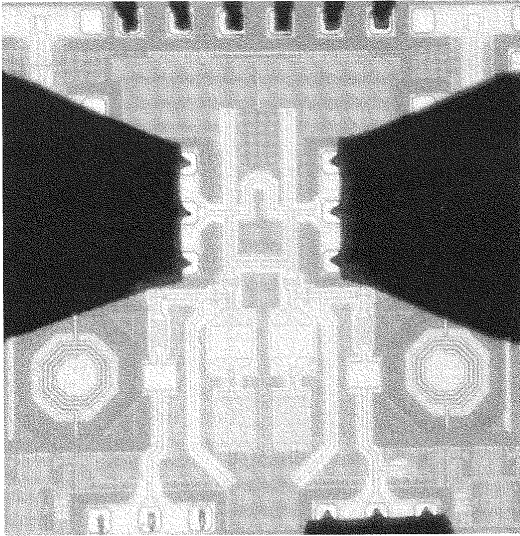


Figure 1.9 A single-transistor 60 GHz CMOS mixer implemented in a digital 130 nm process.

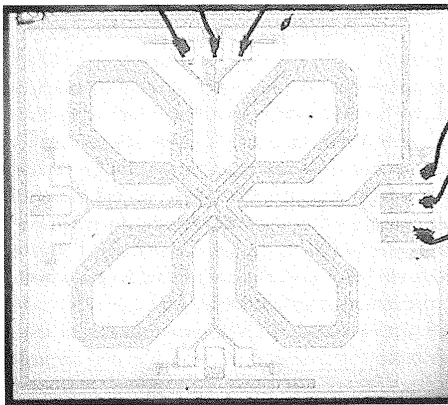


Figure 1.10 A circular standing-wave oscillator operates at 10 GHz [11]. (Copyright 2003, IEEE)

The main bottleneck in optical systems, though, is the electronics. While new optical amplifiers may displace electronic regeneration techniques, routing digital signals on fiber lines requires electronic circuits. Much research effort is dedicated to completely displacing electronic routing through mechanical means. This may seem like a backward step

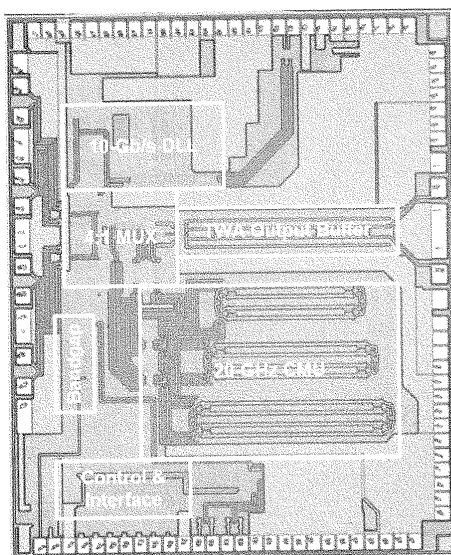


Figure 1.11 A 40 Gb/s SONET MUX/CMU implemented in SiGe technology. The integrated 20 GHz CMU employs coupled oscillators and transmission line resonators [56]. (Copyright 2003, IEEE)

since mechanical systems are inherently much slower than electronic systems. But it is the throughput of electronic systems, not the switching rate, that is limiting optical communication systems. Transistors can only process signals up to a certain limit due to intrinsic f_T and extrinsic parasitics. Micro-electrical-mechanical systems (MEMS) technology allow micromechanical systems to reside on the same substrate as low-cost electronic circuits. A simple MEMS-based hinged mirror switch does not limit the throughput of an optical system, whereas an electronic system requires high-speed digital signals to travel through transistors and parasitics. The co-design of MEMS and electronic systems in a chip requires an understanding of electromagnetics and transduction.

Even if routing of optical signals is done without electronic means, the generation of high-speed optical signals requires electronic systems to pool and un-pool dozens of lower-speed signals on to a fast common serial line. This requires accurate timing and synchronization, such as voltage-controlled oscillators phase locked to reference signals. Optical and high-speed digital circuits require routing of high-speed signals in the presence of board and transmission line parasitics, as well as the design of passive elements for frequency synthesis. An example optical transceiver chip incorporating the multiplexers (MUX) and clock-multiplication unit (CMU) is shown in Fig. 1.11, where transmission lines are used in a coupled oscillator to generate a low-phase noise high-frequency oscillation.

1.2 System in Package (SiP): chip and package co-design

Present design methodology draws a clear line between the chip and the package. This clean interface between the chip and the package simplifies design considerably, since the two can be analyzed independently. As already noted, the package is a major source of parasitic energy coupling in current transceivers and the careful evaluation of the package is a critical step in the design. At higher frequencies, though, the line between the package and chip will blur even further. In fact, the package could be a crucial and beneficial part of the system, as many key passive components could reside in the package. Package and chip co-design will require new design techniques and a more intimate understanding of the electromagnetic effects.

The simultaneous simulation of the interaction of the board environment, the package, and the on-chip passives is presently beyond the scope of present-day simulation tools. One problem is the vastly different scales employed in board environment (centimeters), package (sub-cm), and chip environment (sub-micron geometries). While future tools may overcome this limitation, present-day engineers must have a firm grasp of fundamental engineering to analyze and design such structures. For instance, instead of simulating the entire system, key coupling issues can be dealt with only by simulating critical sub-portions of the system. Identification of key elements and analysis of such systems are two of the goals of this book.

1.3 Future wireless communication systems

The conceptual transceiver shown in Fig. 1.12 naturally partitions into several sections, most notably the RF front-end section, the analog core, and the digital signal processing (DSP) unit. The arrows represent the flow and transfer of information among the various sections. The RF receive section amplifies the weak incoming RF signal and converts it into an intermediate frequency. The choice of IF frequency mostly determines the architecture of the system. The analog section performs further amplification and filtering and converts the signal into digital form. From here, the signal is usually buffered and driven off-chip into the DSP core where the final demodulation and detection are performed. Similarly, on the transmit side, information is encoded and modulated in the DSP core. The analog section converts the signal from digital discrete time to continuous time and up-converts the center frequency to the transmit band. In this transceiver each section communicates with every other section in a well-defined manner. Isolation between the various sections is achieved by physical separation and proper frequency planning to avoid interference.

There are several things to note about this system. Note that the entire system is physically much smaller than the wavelength. Thus, the quasi-static approximation is applicable and much of the internal circuitry can be considered as lumped elements. Also, due to the small size, there is no significant coupling between the external electromagnetic fields propagating in free space and the on-chip signals. The transduction of RF energy from free space to the volume of the conductors is achieved externally through an antenna element. This naturally partitions the radiation incident on the system from the transceiver. The RF energy flows

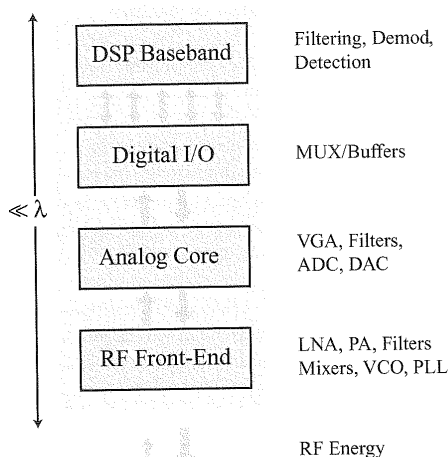


Figure 1.12 A block diagram of a present-day advanced wireless transceiver.

into the system along a well-defined path from the package interface between the board and the chip. Due to the physically large size and requirement for high quality, many passive devices in this system are realized externally, such as surface acoustic wave (SAW) filters, wire-wound high-quality inductors, and low-insertion loss diode switches. To save cost and power, the trend has been to integrate as many of the passive devices on-chip as possible. At higher frequencies this is increasingly a necessity due to the large parasitic impedances present in the board environment. MEMS technology may allow many off-chip components to be miniaturized and moved on chip.

To improve isolation in present-day transceivers, many options such as deep trenches, triple-well devices, or special technologies, such as SOI (Silicon On Insulator), are effective. The proper grounding and regulation of supplies are also important. While these techniques are effective at mitigating the coupling at lower microwave frequencies, they are much less effective at higher frequencies.

A hypothetical transceiver at a much higher frequency is shown in Fig. 1.13. Due to the large physical size of the chip and package relative to the wavelength, an array of antenna elements is integrated directly into the package. The DSP core is the heart of the system and there is constant feedback between the baseband and RF sections to optimize the power and throughput in the system. For instance, the phased antenna array can be used to boost the directivity of the receiver or transmitter to improve the SNR or to lower the power. Spatial diversity in the antenna elements can also boost the data throughput per fixed bandwidth beyond Shannon's limits.

We see that the package now plays a much larger role in this hypothetical transceiver. The package is physically much larger than the chip and contains an ever-increasing density

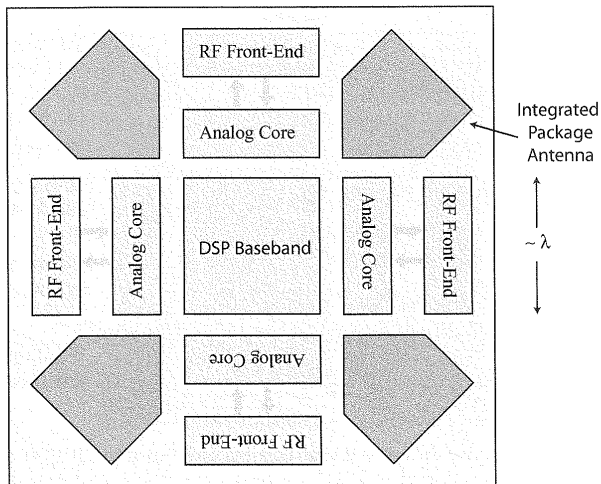


Figure 1.13 A block diagram of a hypothetical future wireless transceiver.

of signal lines in close proximity. If the package design is not performed simultaneously as an integral part of the system, the electromagnetic coupling can serve as a parasitic feedback path that can compromise the stability and lower the sensitivity of the system. These problems are naturally more severe at higher frequencies. An example of a high-frequency system with large levels of microwave integration is shown in Fig. 1.14 [24], where a bank of eight receivers operate in parallel to down-convert signals from an antenna array. The receiver can select the appropriate phase shift in each path to form a high-directivity antenna, or to improve the spatial diversity of the receiver. Proper routing of signals and isolation are key elements in the successful design of such a system.

1.4 Circuits and electromagnetic simulation

Circuits are really a part of electromagnetics. As we alluded to earlier, circuit theory can adequately replace electromagnetics when circuit dimensions are much smaller than the wavelength of electromagnetic fields. This allows all time variation to be discarded and difficult partial differential equations (PDE) resulting from Maxwell's equations to be reduced to ordinary differential equations (ODE).

Many tools are available for solving a system of non-linear ODEs resulting from applying KVL and KCL to nodes and loops in a circuit. Since Maxwell's equations are linear, the non-linearity is introduced primarily through the transistors and junctions in the circuit. The most common tool for solving ODEs in the circuit designer's arsenal is SPICE and its many variants. Frequency domain solvers (harmonic balance) and period steady-state (PSS)

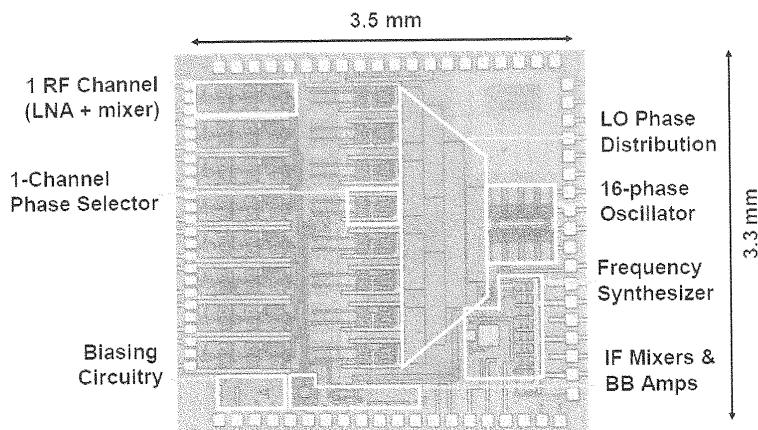


Figure 1.14 A fully integrated SiGe phased array front-end operating at 24 GHz [24]. (Copyright 2003, IEEE)

solvers are also available and are helpful when analyzing periodic time-variant circuits in the presence of noise.

PDE solvers, by contrast, are much slower and harder to use than ODE solvers such as SPICE. This is because a space of three to four dimensions (three spatial dimensions and one dimension for time) is involved. This has a large impact on traditional matrix solvers that require $O[n^2-3]$ time to solve a system of n unknowns. Time variation can be removed for a harmonic steady-state solution, but this even further complicates incorporating PDE solutions in an ODE solver. For instance, if you simulate an inductor with an electromagnetic (EM) solver, how do you take the frequency domain solution into a time-domain solver and efficiently compute the solution?

Electromagnetic simulation and analysis is a critical component in successfully designing active and passive components. This is missing from the circuit curriculum, since circuit designers usually do not deal with the innards of components such as transistors and inductors. Instead, they employ a lumped circuit model that captures the details. For instance, the BSIM models incorporated in Berkeley SPICE and commercial simulators model transistor device physics with hundreds of elements and parameters to describe a sub-micron transistor accurately. A transistor, though, is a very small element and most of the physics fit well into a lumped element model. A typical inductor, by contrast, is relatively simple, containing perhaps three elements to model the inductance, the resistive loss, and capacitance of the structure. This model serves as long as the inductor structure is very small relative to the wavelength. But as the frequency of operation increases, several shortcomings of the model are present. For example, the loss is a complicated function of frequency. Even if we only model skin effect, the resistive loss varies like $\sqrt{f}$. How do we put such a variation in resistance in a time-domain SPICE simulation? There is no easy solution.

In addition to SPICE, then, the circuit designer of today and of the future must be able to understand and model electromagnetic effects. In the past each block was partitioned into a lumped model (designed by device and EM engineers) and the circuit designer could ignore the complicated details. For modern circuits this is no longer possible as even short wires on chip have sufficient parasitic inductance, loss, and capacitance as to impact the performance of circuits. The situation is exacerbated by the increasingly higher densities at which circuits are packed on a single die, and electromagnetic coupling becomes the bottleneck in the achievable performance. As EM solution techniques expand with improvements in computational speed and capacity,⁸ the use of an EM solver will become an integrated part of the design of circuits much as using SPICE is standard practice today.

⁸ It is interesting to note that each generation of advance is fueled by new computation tools running on the previous generation of computers. This self-feeding cycle of advancement means that the current generation of tools is never good enough for the next generation. That's why we need engineers who understand the fundamentals and can partition difficult problems into digestible pieces. That's why engineers will never be replaced by computers!

2 Capacitance

2.1 Electrostatics review

Let's begin with the all important Gauss' Theorem. It is easy to show that the electric "flux" crossing any sphere surrounding a point source is constant and equal to the charge enclosed

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = \int_0^{2\pi} \int_0^\pi \frac{q}{4\pi r^2} r^2 \sin \theta d\theta d\varphi = \frac{q}{4\pi} 2\pi \times 2 = q \quad (2.1)$$

This result is due to the $1/r^2$ dependence of the electric field. Gauss' Law proves that for *any* surface (not just a sphere), the result is identical

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = q \quad (2.2)$$

Furthermore by superposition, the result applies to any distribution of charge

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = \int_V \rho dV = q_{\text{inside}} \quad (2.3)$$

First of all, does it makes sense? For instance, if we consider a region with no net charge, then the flux density crossing the surface is zero. This means that the flux lines entering the surface equal the flux lines leaving the surface. Since charge is the source/sink of electric fields, this does make intuitive sense. We can prove that the flux crossing an infinitesimal surface of any shape is the same as the flux crossing a radial cone. Notice that if the surface $d\mathbf{S}$ is tilted relative to the radial surface by an angle θ , its cross-sectional area is larger by a factor of $1/\cos \theta$. The flux is therefore a constant

$$d\psi = \mathbf{D} \cdot d\mathbf{S} = D_r dS \cos \theta = D_r \frac{dS'}{\cos \theta} \cos \theta = D_r dS' \quad (2.4)$$

For any problem with symmetry, it is easy to calculate the fields directly using Gauss' Law. Take a long (infinite) charged wire as shown in Fig. 2.1. If the charge density is constant with density given by λ C/m, then by symmetry the field is radial. Applying Gauss' Law to a small concentric cylinder surrounding the wire

$$\oint \mathbf{D} \cdot d\mathbf{S} = D_r 2\pi r d\ell \quad (2.5)$$

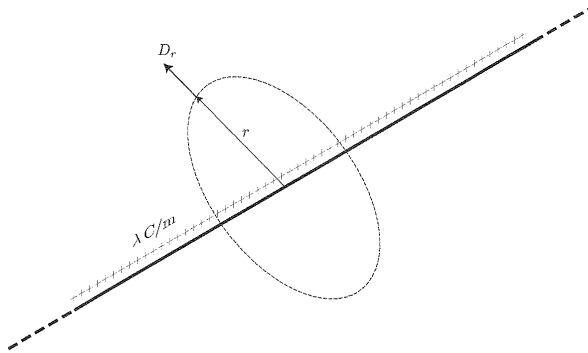


Figure 2.1 An infinitely long charged wire.

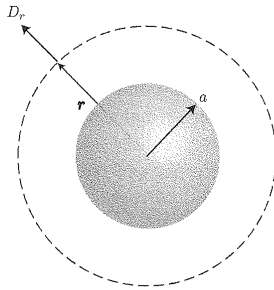


Figure 2.2 A cloud of charge of uniform density.

Since the charge inside the cylinder is simply $\lambda d\ell$

$$D_r 2\pi r d\ell = \lambda d\ell \quad (2.6)$$

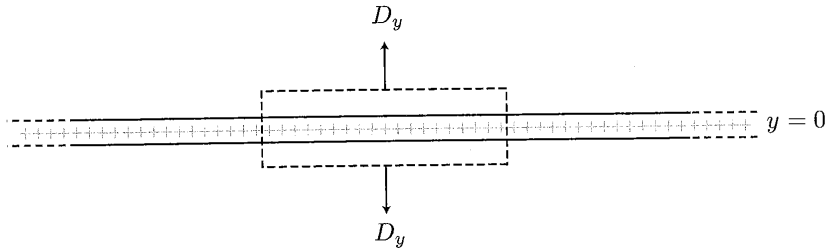
$$D_r = \frac{\lambda}{2\pi r} \quad (2.7)$$

Another easy problem is a cloud of charge Q shown in Fig. 2.2. Since the charge density is uniform $\rho = Q/V$, and $V = \frac{4}{3}\pi a^3$

$$\oint \mathbf{D} \cdot d\mathbf{S} = \begin{cases} \rho \frac{4}{3}\pi r^3 = Q \left(\frac{r}{a}\right)^3 & r < a \\ Q & r > a \end{cases} \quad (2.8)$$

But $\oint \mathbf{D} \cdot d\mathbf{S} = 4\pi r^2 D_r$

$$D_r = \begin{cases} \frac{Qr}{4\pi a^3} & r < a \\ \frac{Q}{4\pi r^2} & r > a \end{cases} \quad (2.9)$$

Figure 2.3 A plane charged uniformly with surface density ρ_s .

Consider an infinite plane that is charged uniformly with surface charge density ρ_s , shown in Fig. 2.3. By symmetry, the flux through the sides of a centered cylinder intersecting with the plane is zero and equal at the top and bottom. The flux crossing the top, for instance, is simply DdS , where D can only possess a $\hat{y}$ component by symmetry. The total flux is thus $2DdS$. Applying Gauss' Law

$$2DdS = \rho_s dS \quad (2.10)$$

The electric field is therefore

$$\mathbf{E} = \hat{y} \begin{cases} \frac{\rho_s}{2\epsilon_0} & y > 0 \\ -\frac{\rho_s}{2\epsilon_0} & y < 0 \end{cases} \quad (2.11)$$

We can reformulate Gauss' Law in differential form as follows. Applying the definition of divergence to the electric flux density D , we have

$$\lim_{V \rightarrow 0} \frac{\oint_S \mathbf{D} \cdot d\mathbf{S}}{V} = \frac{q_{\text{inside}}}{V} = \rho = \nabla \cdot \mathbf{D} \quad (2.12)$$

Therefore we have the important result that at any given point

$$\nabla \cdot \mathbf{D} = \rho(x, y, z) \quad (2.13)$$

This is the analog to the equivalent statement

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = \int_V \rho dV \quad (2.14)$$

The *Divergence Theorem* is a direct proof of this relationship between the volume integral of the divergence and the surface integral (applies to any vector function $\mathbf{A}$)

$$\int_V \nabla \cdot \mathbf{A} dV = \oint_S \mathbf{A} \cdot d\mathbf{S} \quad (2.15)$$

Here V is any bounded volume and S is the closed surface on the boundary of the volume. Application of this theorem to the electric flux density D immediately gives us the differential form of Gauss' Law from the integral form.

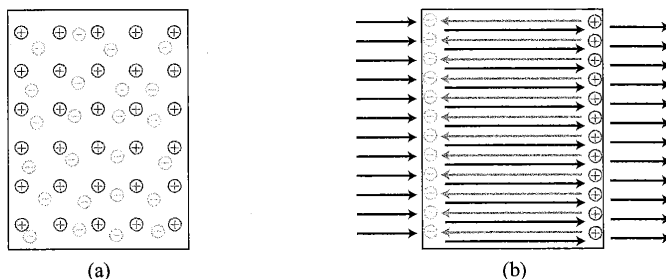


Figure 2.4 (a) An ideal conductor can be modeled as a sea of free electrons roaming about among ionized background molecules. (b) These electrons readily respond to an external field. The rearrangement of charge produces an internal field that perfectly balances (cancels) the applied field.

Perfect conductors

Perfect conductors are idealized materials with zero resistivity. Common metals such as gold, copper, and aluminum are examples of good conductors.

A fuzzy picture of a perfect conductor, shown in Fig. 2.4a, is a material with many mobile charges that easily respond to an external field. These carriers behave as if they are in a vacuum and they can move unimpeded through the metal. These mobile charges, though, cannot leave the material due to a large potential barrier (related to the work function of the material).

We may argue that the electric field is zero under static equilibrium based on the following argument. Under static equilibrium charges cannot move. So we might argue that if an external field is applied, then all the charges in a perfect conductor would move to the boundary where they would remain due to the potential barrier. But this argument ignores the fact that the external field can be canceled by an *internal* field due to the rearrangement of charges (see Fig. 2.4b).

Thus we have to invoke Gauss' Law to further prove that in fact there can be no net charge in the body of a conductor. Because if net charge exists in the body, Gauss' Law applied to a small sphere surrounding the charge would require a field in the body, which we already hypothesized to be zero. Therefore our argument is consistent. In conclusion, we are convinced that a perfect metal should have zero electric field in the body and no net charge in the body. Thus, we expect that if any net charge is to be found in the material, it would have to be on the boundary.

We could in fact define a perfect conductor as a material with zero electric field inside the material. This is an alternative way to define a perfect conductor without making any assumptions about conductivity (which we have not yet really explored). A perfect conductor is also an equipotential material under static conditions, or the potential is everywhere constant on the surface of a perfect conductor. This is easy to prove since if $E \equiv 0$ in the material, then $\int \mathbf{E} \cdot d\mathbf{l}$ is likewise zero between any two points in the material body.

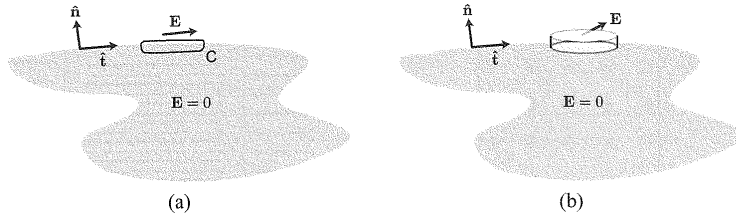


Figure 2.5 (a) The tangential boundary condition can be found by calculating the line integral of the field over a small path partially penetrating the conductor. (b) The normal boundary condition is discovered by calculating the flux of the field through a small cylinder (“pill box”) partially embedded in the conductor.

Perfect conductor boundary conditions

Let us begin with *tangential* boundary conditions, shown in Fig. 2.5a. It is easy to show that the electric field must cross the surface of a perfect conductor at a normal angle. Since we have

$$\int_C \mathbf{E} \cdot d\mathbf{l} \equiv 0 \quad (2.16)$$

for any path C , choose a path that partially crosses into the conductor. Since $E = 0$ inside the conductor, the only contributions to the integral are the side walls and E_t , or the tangential component along the path.

In the limit, we can make the path smaller and smaller until it is tangent to the surface and imperceptibly penetrates the material so the side wall contributions vanish.

$$\int_C \mathbf{E} \cdot d\mathbf{l} = E_t \Delta \ell \equiv 0 \quad (2.17)$$

We are thus led to conclude that $E_t \equiv 0$ at the surface. Another argument is that since $\mathbf{E} = -\nabla\phi$ and furthermore since the surface is an equipotential, then clearly E_t is zero since there can be no change in potential along the surface.

Next consider the *normal* boundary conditions. Consider a “pill-box” hugging the surface of a perfect conductor as shown in Fig. 2.5b. The electric flux density is computed for this surface. As before, we will make the volume smaller and smaller until the sidewall contribution goes to zero. Since the bottom is still in the conductor and the field is zero, this term will not contribute to the integral either. The only remaining contribution is the normal component of the top surface

$$\oint \mathbf{D} \cdot d\mathbf{S} = D_n dS = Q_{\text{inside}} = dS \rho_s \quad (2.18)$$

We have thus shown that $D_n = \rho_s$

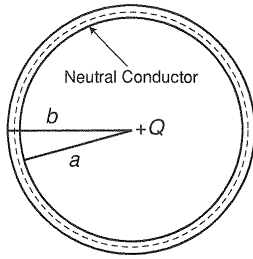


Figure 2.6 Charge placed inside a neutral conducting spherical shell.

Example 1

Consider a charge placed at the center of a spherical shell made of perfect conducting materials. Let us find the surface charge density on the inner and outer surface of the conductor.

Let us apply Gauss' Law for a sphere lying inside the conductor (shown with the dashed line in Fig. 2.6). Since $\mathbf{E} \equiv 0$ on this surface, the charge inside must likewise be zero $Q_{\text{inside}} = 0$, which implies that there exists a uniform charge density (by symmetry) of $\rho_{\text{inner}} = -Q/S_i$ where $S_i = 4\pi a^2$.

If we now consider a larger sphere of radius $r > b$, since the sphere is neutral, the net charge is just the isolated charge Q . Thus

$$D_r 4\pi r^2 = Q + \underbrace{\rho_{\text{inside}} S_1}_{-Q} + \rho_{\text{outside}} S_2 = Q \quad (2.19)$$

Thus $\rho_{\text{outside}} = Q/S_2 = \frac{Q}{4\pi b^2}$

The fields have been sketched in Fig. 2.7. Radial symmetry is preserved and an induced negative and positive charge density appear at the shell surfaces.

In summary, the radial electric fields are given by (and plotted in Fig. 2.8)

$$\begin{aligned} D_r &= \frac{Q}{4\pi r^2} & r < a \\ D_r &= 0 & a \leq r \leq b \\ D_r &= \frac{Q}{4\pi r^2} & r > b \end{aligned} \quad (2.20)$$

Let us compute the potential from the fields. Take the reference point at ∞ to be zero and integrate along a radial path. For $r > b$

$$\phi = - \int_{\infty}^r E_r dr = \frac{Q}{4\pi \epsilon r} \quad (2.21)$$

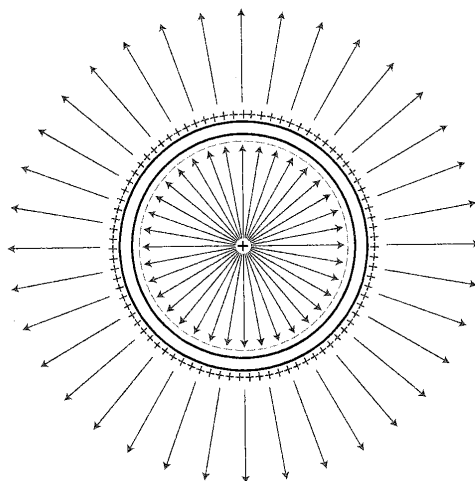


Figure 2.7 Induced charge and electric fields on a neutral spherical shell.

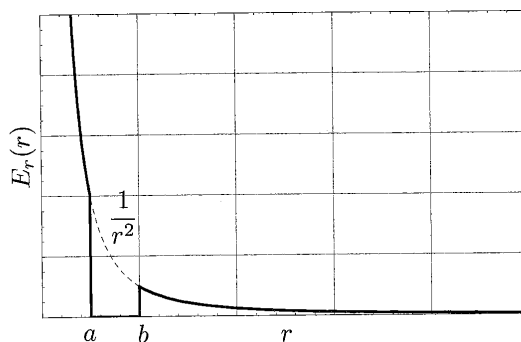


Figure 2.8 Plot of radial electric flux density for the neutral spherical shell.

Likewise for $a \leq r \leq b$, since E_r is zero for the path inside the conducting sphere, the potential remains constant at $\frac{Q}{4\pi\epsilon_0 b}$ until we exit the conductor. Finally, once outside the conductor for $r > b$ we continue integrating

$$\phi = \frac{Q}{4\pi\epsilon_0 b} - \int_a^r \frac{Q}{4\pi\epsilon_0 r^2} dr \quad (2.22)$$

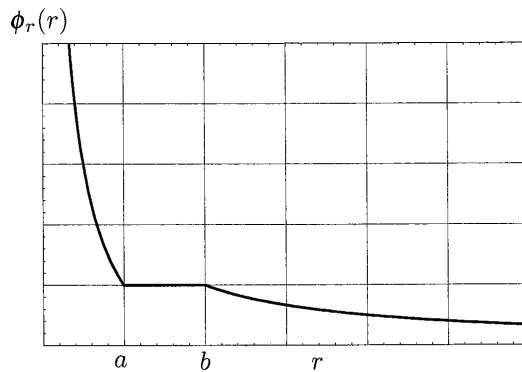


Figure 2.9 The electric potential for a charge placed at the center of a spherical shell.

Thus we have

$$\phi = \frac{Q}{4\pi\epsilon_0} \left(\frac{1}{b} + \frac{1}{r} - \frac{1}{a} \right) \quad (2.23)$$

A plot of the potential is shown in Fig. 2.9. Note that potential on the conductor is non-zero and its value depends on the amount of charge at the center.

Example 2

What if we now ground the shell from the previous example? That means that $\phi = 0$ on the surface of the conductor. The work done in moving a point charge from infinity to the surface of the shell is thus zero. Choose a radial path

$$\phi = - \int_{\infty}^r E_r dr = 0 \quad (2.24)$$

Since the function E_r is monotonic, $E_r = 0$ everywhere outside the sphere! A grounded spherical shell acts like a good shield. If $E_r(b) = 0$, then $\rho_s = 0$ at the outer surface as well.

To find the charge on the inner sphere, we can use the same argument as before unchanged so $\rho_{\text{inside}} = -Q/S_1$. But now the material is not neutral! Where did the charge go? Imagine starting with the ungrounded case where the positive charge is induced on the outer surface. Then ground the sphere and notice that charge flows out of the sphere into ground thus “charging” the material negatively.

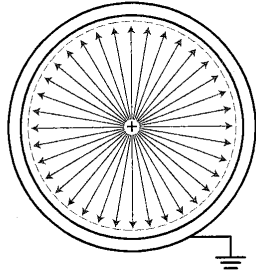


Figure 2.10 Induced charge and electric fields on a grounded spherical shell.

The fields have been sketched in Fig. 2.10. Radial symmetry is preserved and an induced negative charge density appears at inner shell surfaces. The field outside the shell is zero.

Poisson's equation (aka Fish equation)

We already have an equation for calculating $\mathbf{E}$ and ϕ directly from a charge distribution and vice versa. If we know $\mathbf{E}$, then we can find ρ by simply taking the divergence, $\rho = \epsilon \nabla \cdot \mathbf{E}$. Alternatively if we know ϕ

$$\rho = \epsilon \nabla \cdot \mathbf{E} = -\epsilon \nabla \cdot \nabla \phi \quad (2.25)$$

The operator $\nabla \cdot \nabla$ is a new operator and we call it the Laplacian ∇^2 . Thus we have Poisson's equation $\nabla^2 \phi = -\frac{\rho}{\epsilon}$. In a charge-free region we have $\nabla^2 \phi = 0$. This is known as Laplace's equation. Let us untangle this beast of an operator

$$\nabla \cdot \nabla \phi = \left(\frac{\partial}{\partial x} \hat{\mathbf{x}} + \frac{\partial}{\partial y} \hat{\mathbf{y}} + \frac{\partial}{\partial z} \hat{\mathbf{z}} \right) \cdot \left(\frac{\partial \phi}{\partial x} \hat{\mathbf{x}} + \frac{\partial \phi}{\partial y} \hat{\mathbf{y}} + \frac{\partial \phi}{\partial z} \hat{\mathbf{z}} \right) \quad (2.26)$$

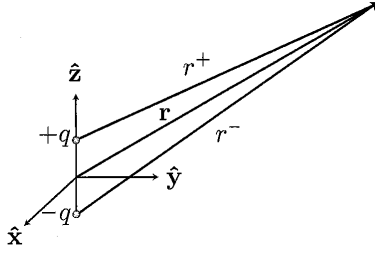
since all cross products involving mutual vectors die (e.g. $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} \equiv 0$), we have

$$\nabla^2 \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} \quad (2.27)$$

Other coordinate systems are not so nice!

Dielectrics

So far we have only considered electrostatics in a vacuum and with ideal conductors. We now expand our discussion to include dielectric materials. Dielectric materials are often good insulators. A perfect insulator cannot conduct DC current because there are no mobile charges as all electrons (free carriers) in the material are bound to the nuclei. But there is a response to an external field due to the dielectric properties of the material.

Figure 2.11 Calculation of the electric potential at a point $\mathbf{r}$ distant from the dipole.

Review of dipoles

An electric dipole consists of two equal and opposite charges, q , separated by a distance, d . With reference to Fig. 2.11, consider the potential due to an electric dipole at points far removed from the dipole $r \gg d$

$$\phi(\mathbf{r}) = \frac{q}{4\pi\epsilon r^+} + \frac{-q}{4\pi\epsilon r^-} = \frac{q}{4\pi\epsilon} \left(\frac{1}{r^+} - \frac{1}{r^-} \right) \quad (2.28)$$

We can simplify the above expression by the application of the Law of Cosines, since vector summation is fundamentally related to a triangle. Note we can think of adding two vectors $\mathbf{r}$ and $\frac{d}{2}\hat{\mathbf{z}}$ as forming $\mathbf{r}^+ = \mathbf{r} + \frac{d}{2}\hat{\mathbf{z}}$. The length of a vector is given by $|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a}$. We can therefore express r^+ in terms of r and d

$$r^+ = \sqrt{\mathbf{r}^+ \cdot \mathbf{r}^+} = \sqrt{\left(\mathbf{r} + \frac{d}{2}\hat{\mathbf{z}}\right) \cdot \left(\mathbf{r} + \frac{d}{2}\hat{\mathbf{z}}\right)} \quad (2.29)$$

or simplifying a bit and recalling that $r \gg d$

$$r^+ = r \sqrt{1 + \left(\frac{d}{2r}\right)^2 + \frac{d}{r} \cos \theta} \approx r \sqrt{1 + \frac{d}{r} \cos \theta} \quad (2.30)$$

So for the potential we have

$$\frac{1}{r^+} = \frac{1}{r} \frac{1}{\sqrt{1 + \frac{d}{r} \cos \theta}} \approx \frac{1}{r} \left(1 - \frac{d}{2r} \cos \theta \right) \quad (2.31)$$

and

$$\frac{1}{r^-} \approx \frac{1}{r} \left(1 + \frac{d}{2r} \cos \theta \right) \quad (2.32)$$

So that the potential is finally given

$$\phi(\mathbf{r}) = \frac{q}{4\pi\epsilon} \left(\frac{1}{r^+} - \frac{1}{r^-} \right) = \frac{-qd \cos \theta}{4\pi\epsilon r^2} \quad (2.33)$$

Unlike an isolated point charge, the potential drops like $1/r^2$ rather than $1/r$.

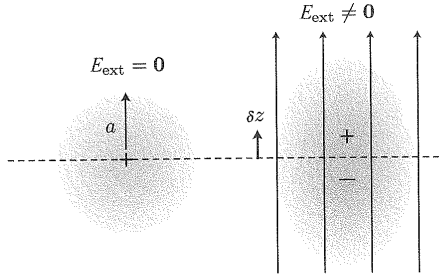


Figure 2.12 A spherically symmetric electron cloud has zero average dipole moment. But under the application of an external field, E_{ext} , the electron cloud is distorted (exaggerated for clarity) to produce a dipole moment.

The potential calculation was easy since we did not have to deal with vectors. The electric field is simply $\mathbf{E} = -\nabla\phi$. Since our answer is in spherical coordinates, and there is no ϕ variation due to symmetry, we have

$$\nabla\phi = \hat{\mathbf{r}} \frac{\partial\phi}{\partial r} + \hat{\boldsymbol{\theta}} \frac{1}{r} \frac{\partial\phi}{\partial\theta} = \hat{\mathbf{r}} \frac{2qd \cos\theta}{4\pi\epsilon r^3} + \hat{\boldsymbol{\theta}} \frac{1}{r} \frac{qd}{4\pi\epsilon r^2} \sin\theta \quad (2.34)$$

Thus the electric field is given by

$$\mathbf{E} = -\frac{qd}{4\pi\epsilon r^3} (\hat{\mathbf{r}} 2 \cos\theta + \hat{\boldsymbol{\theta}} \sin\theta) \quad (2.35)$$

We can define a vector dipole moment $\mathbf{p} = q\mathbf{d}$, where $\mathbf{d}$ is a vector connecting the positive and negative charge. In terms of $\mathbf{p}$, we can rewrite the potential in a coordinate independent manner

$$\phi(\mathbf{r}) = \frac{\hat{\mathbf{r}} \cdot \mathbf{p}}{4\pi\epsilon r^2} \quad (2.36)$$

For a spherically symmetric atom and electron cloud distribution, shown in Fig. 2.12, the time average dipole moment is zero. But the application of an external field can distort the electronic charge distribution to produce a net dipole.

It is easy to see why this happens. The electrons are bound to the nucleus through the internal force of the atom. This force is much stronger than our applied field and so the probability that an electron will obtain sufficient energy from the field to leave the atom is negligible. But the average electronic distribution is the minimum energy configuration. If the electrons displace slightly from their equilibrium positions, then the new configuration yields lower overall energy in the presence of a field.

It is important to note the displacement $\Delta z \ll a$. This is because the *internal* field of an atom is already very strong

$$E_{\text{int}} \sim \frac{e}{4\pi\epsilon a^2} \sim 10^{11} \text{ V/m} \quad (2.37)$$

This is several orders of magnitude larger than any field we typically impart on to the atom externally. To first order, then, the response is linear with the field. This is because for a

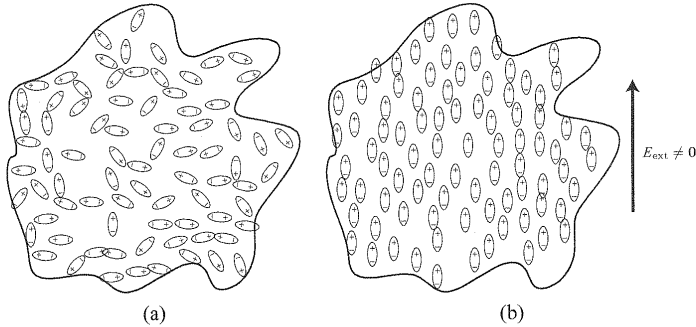


Figure 2.13 (a) A material consisting of molecules with net dipole moments has a net zero dipole moment due to the random orientation of the molecules. (b) The dipoles tend to align under the application of a sufficiently strong electric field.

small displacement, any function looks linear

$$f(x + \Delta x) \approx f(x) + \Delta x f'(x) \quad (2.38)$$

Atomic dipole moment

Based on the fact that $E_{\text{ext}} \ll E_{\text{int}}$, we expect that the induced dipole moment in the atom to be small

$$p = e\Delta z \quad (2.39)$$

But since there are a lot of atoms, the net effect can be large. The constant of proportionality is known as the electronic polarizability coefficient α_e

$$\mathbf{p} = \alpha_e \mathbf{E} \quad (2.40)$$

Typically $\alpha_e \sim 10^{-40} \text{F} \cdot \text{m}^2$.

Ionic polarizability

For more complex molecules, like CH_4 , there can be larger-scale polarization. Instead of each individual atom getting polarized, the shared covalent bond electrons can rearrange and produce a larger polarization. In other cases, the molecules may have built-in polarization due to ionic or covalent bonds. Water is a good example of this. As shown in Fig. 2.13a, normally the dipole moments are pointing in random directions so the net polarization for $E_{\text{ext}} = 0$. But the application of an external field can align the dipoles, shown in Fig. 2.13b. The response can be huge due to the large number of molecules involved.

Unfortunately, the effect of thermal energy tends to re-randomize the orientation so we expect that on average only a small fraction of the dipoles are aligned. In fact, we can show that the probability that a molecule be aligned is related to the energy of a dipole in an

external field. The energy is proportional to p^2 and so the fraction of polarized molecules depends on the ratio of p^2 and $3k_B T$

$$\alpha_0 = \frac{p^2}{3k_B T} \quad (2.41)$$

Typically $\alpha_0 \sim 10^{-38} \text{Fm}^2$

Total polarization

In general, the total polarization arises from different sources (induced and permanent). If the density of polarization is N per volume, the net polarization is

$$\mathbf{P} = N\alpha\mathbf{E} = \epsilon_0\chi_e\mathbf{E} \quad (2.42)$$

χ_e is known as the electric susceptibility. We are now in a position to derive the potential due to polarized matter. Let us rewrite the potential due to a single dipole

$$\phi(r) = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p} \cdot \hat{\mathbf{r}}}{r^2} = -\frac{1}{4\pi\epsilon_0} \mathbf{p} \cdot \nabla \frac{1}{r} \quad (2.43)$$

Notice that $\nabla \frac{1}{r} = \nabla r^{-1} = \hat{\mathbf{r}} \frac{\partial r^{-1}}{\partial r} = -\frac{\hat{\mathbf{r}}}{r^2}$. Integrating over all dipoles we have

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int_V -\mathbf{P} \cdot \nabla \frac{1}{|\mathbf{r} - \mathbf{r}'|} dV \quad (2.44)$$

Note that $\nabla R^{-1} = -\nabla' R^{-1}$, where $R = |\mathbf{r} - \mathbf{r}'|$. Using the identity $\nabla \cdot (a\mathbf{A}) = a\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla a$

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \left(\int_V \nabla' \cdot \frac{\mathbf{P}}{|\mathbf{r} - \mathbf{r}'|} dV - \int_V \frac{\nabla' \cdot \mathbf{P}}{|\mathbf{r} - \mathbf{r}'|} dV \right) \quad (2.45)$$

By Divergence Theorem we can recast this into

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \left(\oint_S \frac{\mathbf{P} \cdot \hat{\mathbf{n}}}{|\mathbf{r} - \mathbf{r}'|} dS - \int_V \frac{\nabla' \cdot \mathbf{P}}{|\mathbf{r} - \mathbf{r}'|} dV \right) \quad (2.46)$$

Effective volume and surface charge

Notice that these integrals look like an ordinary potential calculation for surface charge and volume charge. Define a surface charge density $\rho_s = \mathbf{P} \cdot \hat{\mathbf{n}}$ and a volume charge density $\rho_v = -\nabla' \cdot \mathbf{P}$. Then we have

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \left(\oint_S \frac{\rho_s}{|\mathbf{r} - \mathbf{r}'|} dS + \int_V \frac{\rho_v}{|\mathbf{r} - \mathbf{r}'|} dV \right) \quad (2.47)$$

There must be a good explanation for this coincidence! Here is the intuitive picture. If the material is uniform and the external field is uniform, then the polarization is also uniform and $\nabla' \cdot \mathbf{P} = 0$. The interpretation is that all the internally induced polarizations cancel out.

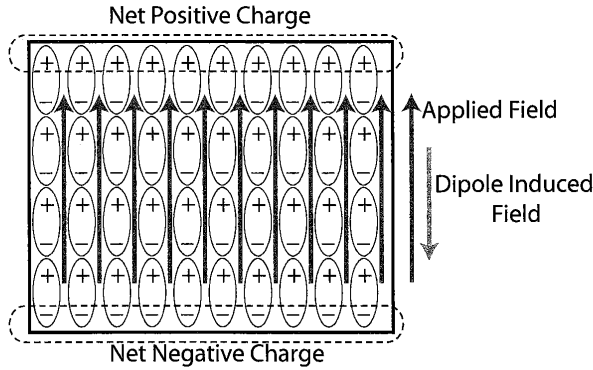


Figure 2.14 For a uniform material polarized by an external field, the internal dipole moments cancel. Only the dipoles on the boundary produce net charge.

As shown in Fig. 2.14, the dipoles on the boundary, though, remain unbalanced and thus produce a net charge. If the polarization is non-uniform, though, then $\nabla' \cdot \mathbf{P} \neq 0$ and a volume polarization arises.

Now back to Gauss' Law. Since the bound charges can create net charge, we must include their contribution in Gauss' Law

$$\nabla \cdot \epsilon_0 \mathbf{E} = \rho_{\text{free}} + \rho_{\text{bound}} \quad (2.48)$$

where $\rho_{\text{bound}} = -\nabla \cdot \mathbf{P}$. Substituting in the above equation

$$\nabla \cdot \epsilon_0 \mathbf{E} + \nabla \cdot \mathbf{P} = \rho_{\text{free}} \quad (2.49)$$

Grouping terms we have (where ρ is now understood to be free of charge)

$$\nabla \cdot (\epsilon_0 \mathbf{E} + \mathbf{P}) = \rho_{\text{free}} = \rho \quad (2.50)$$

Electric flux density vector

We can re-define the electric flux density

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \quad (2.51)$$

so that it depends only on free charge

$$\nabla \cdot \mathbf{D} = \rho \quad (2.52)$$

Since the polarization is proportional to $\mathbf{E}$

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \epsilon_0 \chi_e \mathbf{E} = \epsilon_0 (1 + \chi_e) \mathbf{E} \quad (2.53)$$

The dielectric constant captures the polarization effects in a macroscopic coefficient $\epsilon = \epsilon_r \epsilon_0 = \epsilon_0 (1 + \chi_e)$

Dielectric boundary conditions**Normal D**

Applying Gauss' Law to the interface of two materials

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = (\hat{\mathbf{n}} \cdot \mathbf{D}_1) \Delta S - (\hat{\mathbf{n}} \cdot \mathbf{D}_2) \Delta S = \rho_s \Delta S \quad (2.54)$$

Thus the discontinuity in the normal field must be due to surface charge

$$\mathbf{D}_{1n} - \mathbf{D}_{2n} = \rho_s \quad (2.55)$$

For dielectrics, there is no "free" surface charge and so $\rho_s = 0$

$$\hat{\mathbf{n}} \cdot \mathbf{D}_1 = \hat{\mathbf{n}} \cdot \mathbf{D}_2 \quad (2.56)$$

or $D_{1n} = D_{2n}$. The normal electric field is discontinuous $\epsilon_1 E_{1n} = \epsilon_2 E_{2n}$.

Tangential E

Take the line integral around the interface. Since the field is static, the line integral along any closed path is identically zero

$$\int_C \mathbf{E} \cdot d\mathbf{l} = (E_{1t} - E_{2t}) \Delta w = 0 \quad (2.57)$$

This means that the tangential field is always continuous, $E_{1t} = E_{2t}$. Then the tangential component of D is *not* continuous

$$\frac{D_{1t}}{\epsilon_1} = \frac{D_{2t}}{\epsilon_2} \quad (2.58)$$

2.2 Capacitance

Capacitance of a conductor is defined as

$$C \triangleq \frac{q}{\phi} \quad (2.59)$$

where a larger C means that an object can store more charge at a fixed potential. It is not so different than the volume of a tank of water. The bigger the tank, the more water it can store. Hence it is a measure of capacity to store charge. As an example, consider an isolated spherical conductor of radius a . By Gauss' Law we know the radial field is given by ($r \geq a$)

$$D_r = \frac{q}{4\pi r^2} \quad (2.60)$$

If we integrate this field to arrive at the potential, we have

$$\phi = - \int_{\infty}^a E_r dr = - \frac{q}{4\pi\epsilon} \int_{\infty}^a \frac{dr}{r^2} = \frac{q}{4\pi\epsilon} \left(\frac{1}{r} \right)_{\infty}^a = \frac{q}{4\pi\epsilon a} \quad (2.61)$$

The capacitance is therefore

$$C = 4\pi\epsilon a \propto a \quad (2.62)$$

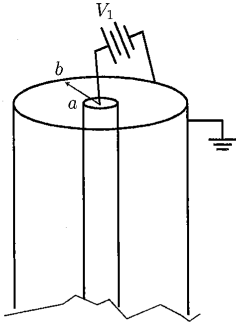


Figure 2.15 Setup for the calculation of the capacitance per unit length of a coaxial structure.

The larger sphere can hold more charge at a fixed potential. Equivalently, dumping a given charge on to a larger sphere will increase its potential less than a smaller sphere.

We can also define the capacitance between two objects as

$$C_{12} = \frac{q_1}{\phi_{12}} \quad (2.63)$$

where we fix the potential between the two objects at ϕ_{12} and measure the amount of charge transfer between the objects. In the above equation a charge of q_1 has been transferred from object 2 to 1 and therefore a charge of $q_2 = -q_1$ will reside on object 2. We can re-interpret the capacitance of a single object as the capacitance relative to a reference at infinity.

Example 3

Let us practice and calculate the capacitance per unit length for a coaxial structure shown in Fig. 2.15. We can solve Poisson's equation to find the potential in the region between the conductors¹

$$\nabla^2 \phi = 0 \quad (2.64)$$

By symmetry ϕ is only a function of r and not a function of θ or z

$$\nabla \cdot \nabla \phi = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \phi}{\partial r} \right) = 0 \quad (2.65)$$

$$r \frac{\partial \phi}{\partial r} = K_1 \quad (2.66)$$

$$\frac{\partial \phi}{\partial r} = \frac{K_1}{r} \quad (2.67)$$

The general solution is therefore $\phi(r) = K_1 \ln r + K_2$.

¹ Note the form of the Laplacian in cylindrical coordinates differs from rectangular coordinates.

This solution must satisfy the boundary conditions that $\phi(a) = V_1$ and $\phi(b) = 0$

$$\phi(r = a) = V_1 = K_1 \ln a + K_2 \quad (2.68)$$

$$\phi(r = b) = 0 = K_1 \ln b + K_2 \quad (2.69)$$

Solving these equations we have

$$\phi(r) = \frac{V_1 \ln r/b}{\ln a/b} \quad (2.70)$$

The field is given by $\mathbf{E} = \hat{\mathbf{r}} E_r = -\hat{\mathbf{r}} \frac{\partial \phi}{\partial r}$

$$\mathbf{E} = \frac{\hat{\mathbf{r}} V_1}{r \ln b/a} \quad (2.71)$$

Recall that $\mathbf{D} \cdot \hat{\mathbf{n}} = \rho_s$ on the surface of conductors. Therefore for a length of ℓ of the coaxial conductors

$$\frac{\epsilon V_1}{a \ln(b/a)} = \frac{q}{2\pi a \ell} \quad (2.72)$$

Solve for the charge to find the capacitance

$$q = \frac{2\pi\epsilon}{\ln \frac{b}{a}} \ell V_1 \quad (2.73)$$

The capacitance per unit length is therefore

$$C' = \frac{C}{\ell} = \frac{2\pi\epsilon}{\ln \frac{b}{a}} \quad (2.74)$$

Parallel plate capacitor

A very convenient idealization is a parallel plate capacitor. Here we take two conducting plates of area A and separate them by a distance t . If $t \ll \min(dx, dy)$, then the field is largely independent of x and y and only a function of z . This is true especially in regions inside the plates far from the edges of the capacitor. Now charge the upper plate to a voltage of V_0 relative to the bottom plate. For convenience take the reference of zero potential at the bottom plate.

If we neglect the edge effects, then the solution of the potential in the charge-free region between the plates is governed by a one-dimensional differential equation.

$$\nabla^2 \phi = \frac{\partial^2 \phi}{\partial z^2} = 0 \quad (2.75)$$

subject to the boundary conditions $\phi(0) = 0$ and $\phi(t) = V_0$. The solution is clearly a linear function (integrate the above equation twice) $\phi(z) = C_1 z + C_2$ and applying the boundary conditions to solve for C_1 and C_2 we arrive at

$$\phi(z) = \frac{V_0}{t} z \quad (2.76)$$

The electric field $E = -\nabla\phi$ can thus be computed by a simple derivative and is a constant within the capacitor

$$\mathbf{E} = -\hat{\mathbf{z}} \frac{V_0}{t} \quad (2.77)$$

By Gauss' Law, the normal component of the electric flux density is equal to the charge density on the plates

$$\mathbf{D} \cdot \hat{\mathbf{n}} = \rho \quad (2.78)$$

Since the field is in fact everywhere perpendicular to the conductor, we simply have

$$\rho_{\text{top}} = -\hat{\mathbf{z}} \cdot \mathbf{D} = \epsilon \frac{V_0}{t} \quad (2.79)$$

and

$$\rho_{\text{bot}} = \hat{\mathbf{z}} \cdot \mathbf{D} = -\epsilon \frac{V_0}{t} \quad (2.80)$$

To find the capacitance we simply note that we are interested in the coefficient $q = CV_0$. To get the total charge, we multiply the constant charge density by the area of the plate $q = \rho_{\text{top}} A = \frac{\epsilon A}{t} V_0 = C V_0$, so the capacitance is given by $C = \frac{\epsilon A}{t}$. This equation is intuitively satisfying. The capacitance goes up with A , since, for a fixed charge on the plates, the charge density drops and so does the potential giving a larger capacitance. Likewise, if we increase t , the capacitance drops since now there is less motivation for positive (negative) charge to flow on to the top (bottom) plate! The charges are more distant from their beloved negative (positive) charges.

For two conductors of any shape, the capacitance is defined as

$$C \triangleq \frac{Q}{\phi_{12}} \quad (2.81)$$

The potential difference ϕ_{12} is the line integral of $\mathbf{E}$ over *any* path from conductor 1 to conductor 2. By Gauss' Law, the positive charge on conductor 1 is equal to the electric flux crossing *any* surface enclosing the conductor. The capacitance is therefore written as

$$C = \frac{\oint_S \mathbf{D} \cdot d\mathbf{S}}{-\int_\Gamma \mathbf{E} \cdot d\mathbf{l}} \quad (2.82)$$

where S surrounds the top plate and Γ is a path from one plate to the other.

From circuit theory we know how to add capacitors in series or in parallel. Parallel caps are easier. If we connect two conductors and connect them to a potential V_0 , the charge is simply the total charge

$$Q = Q_1 + Q_2 = C_1 V_0 + C_2 V_0 = (C_1 + C_2) V_0 \quad (2.83)$$

In general we have

$$C_{||} = C_1 + C_2 + C_3 + \dots \quad (2.84)$$

For series capacitors, note that the applied voltage is divided between the capacitors

$$V_0 = V_1 + V_2 = \frac{Q_1}{C_1} + \frac{Q_2}{C_2} \quad (2.85)$$

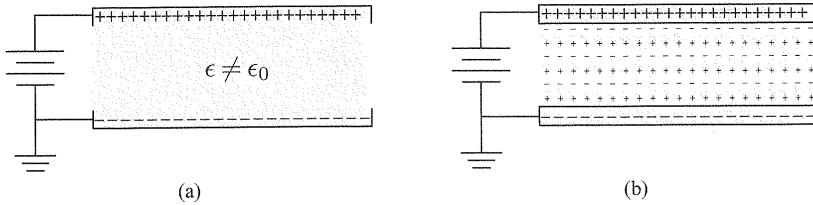


Figure 2.16 A parallel plate capacitor with a dielectric has larger capacitance since the induced dipole moments create an internal electric field which opposes the applied field.

The central observation is that the charge on each capacitor is the same, $|Q_1| = |Q_2|$. This is because the floating node must have zero net charge and thus $-Q_1 = Q_2$. The result can be easily generalized

$$\frac{1}{C_{\text{series}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots \quad (2.86)$$

Parallel plate capacitor with dielectric

One important application for a dielectric is to insulate conductors from one another (coaxial transmission line) while providing mechanical support. In a capacitor there is the additional benefit that the capacitance increases. As shown in Fig. 2.16, the capacitance increases because the effect of the polarization is to produce a net charge density at the top and bottom of the plates that subtracts from the net charge on the conductors.

So for a fixed applied field (voltage), the parallel plate capacitor can support more charge with a dielectric. Another perspective is if we fix the amount of charge on the plates, then the field in the dielectric is reduced and hence the applied electric field is smaller.

Capacitance matrix

The concept of capacitance can be generalized to multiple conductors

$$\begin{aligned} q_1 &= C_{11}V_1 + C_{12}V_2 + C_{13}V_3 + \dots \\ q_2 &= C_{21}V_1 + C_{22}V_2 + C_{23}V_3 + \dots \\ &\vdots \end{aligned} \quad (2.87)$$

Each coefficient C_{ii} represents the self capacitance. It can be computed by applying $V_i = 1\text{V}$ to conductor i while grounding all other conductors. Then C_{ii} is simply the total charge on the conductor. Likewise, to find C_{ij} , we apply a voltage of $V_j = 1\text{V}$ to conductor j while grounding all other conductors. Then C_{ij} is again simply the total charge on conductor i .

When we say we “grounding” all other conductors, we mean that we connect them to a voltage source of zero volts. Recall that voltage is always defined relative to a reference. For instance, we may take one of the conductors as the “ground” reference and then measure all absolute voltages relative to this conductor. Otherwise we may also connect the conductors

to a much larger body, one with infinite capacitance. Then charge can be freely removed or added to the “ground” without changing its potential. Think of using the ocean as the reference “ground.” Then removing or adding several buckets of water will not alter the “level” (voltage) of the ocean.

We may also express the voltage on each conductor in terms of the total charge on each conductor in the system in the following manner

$$\begin{aligned} v_1 &= P_{11}q_1 + P_{12}q_2 + P_{13}q_3 + \dots \\ v_2 &= P_{21}q_1 + P_{22}q_2 + P_{23}q_3 + \dots \\ &\vdots \end{aligned} \quad (2.88)$$

To find P_{ij} , we add a charge of 1C to conductor j and leave all other conductors neutral while observing the voltage at conductor i .

The definition of capacitance follows directly from the laws of electrostatics. Let us place N conductors in vacuum. Then the electric potential at any point in space is given by

$$\phi(r) = \int_V \frac{\rho(r')}{4\pi\epsilon|r-r'|} dV' = \sum_i \int_{V_i} \frac{\rho(r')}{4\pi\epsilon|r-r'|} dV' \quad (2.89)$$

Notice that this equation scales linearly with the absolute amount of charge on conductor i . So we may perform the integral based solely on geometric calculations to obtain coefficient P_{ij}

$$\phi(r) = \sum_i q_i \int_{V_i} \frac{\rho(r')/q_i}{4\pi\epsilon|r-r'|} dV' \quad (2.90)$$

or in matrix form, we may write $v = Pq$. If the matrix P is not singular, we may invert this equation to obtain $q = P^{-1}v$. We may be tempted to call P_{ij}^{-1} a capacitance, but notice that these coefficients are in terms of the potential V_i relative to a common reference

$$\begin{aligned} q_1 &= c_{11}V_1 + c_{12}V_2 + c_{13}V_3 + \dots \\ q_2 &= c_{21}V_1 + c_{22}V_2 + c_{23}V_3 + \dots \\ &\vdots \end{aligned} \quad (2.91)$$

To relate c_{ij} to C_{ij} , simply equate the total charges

$$q_1 = C_{11}V_1 + C_{12}V_2 + C_{13}V_3 = c_{11}V_1 + c_{12}V_2 + c_{13}V_3 + \dots \quad (2.92)$$

Since $V_{1i} = V_1 - V_j$, we have

$$C_{11} = c_{11} + c_{12} + c_{13} + \dots \quad (2.93)$$

Also $C_{ij} = -c_{ij}$, and so these matrices are related but not the same. Notice that $c_{ij} < 0$ is logical, since if we put a positive voltage on node j and observe the induced charge on node i , it should be negative. On the other hand $C_{ij} > 0$, since if we connect the positive terminal of a battery to node i and the negative node of the battery to node j , then the charge on node i should be positive. Capacitors in SPICE are always of the C_{ij} form, and hence positive.

Electrostatic energy of a capacitor

Consider the energy required to charge a capacitor. The amount of work released to move a charge dQ from the positive terminal to the negative terminal is $dU = VdQ$. This work must be stored in the capacitor

$$dU = VdQ = CVdV \quad (2.94)$$

Integrating over the voltage on the capacitor we have

$$U = \frac{1}{2}CV^2 \quad (2.95)$$

We say that this energy is stored in the field of the capacitor. This “field” business will become clear as we develop these ideas further. For a parallel plate capacitor, the field is constant and equal to $E = V/d$. If we substitute in the energy equation, we have

$$U = \frac{1}{2}CE^2d^2 \quad (2.96)$$

Substituting further with $C = \epsilon A/d$

$$U = \frac{1}{2} \frac{\epsilon A}{d} E^2 d^2 = \frac{1}{2} \epsilon \mathcal{V} E^2 \quad (2.97)$$

where $\mathcal{V} = d \times A$ is the volume of the region in between the plates. Since the fields are confined to this volume, we may speculate that the energy density is also so confined.

In this particular case, we have

$$u = \frac{U}{\mathcal{V}} = \frac{1}{2} \epsilon E^2 = \frac{1}{2} \mathbf{D} \cdot \mathbf{E} \quad (2.98)$$

We will show later that this is true in general for any electrostatic field.

Now consider the energy required to charge two conductors to voltages ϕ_1 and ϕ_2 . Let us do the calculation in phases. First apply voltage ϕ_1 to conductor 1 but keep conductor 2 grounded. A charge $q_1 = c_{11}\phi_1$ flows on to conductor 1, whereas a charge $q_2 = c_{21}\phi_1$ flows on to conductor 2. Since conductor 2 is grounded, there is no energy required to add or remove charge from it. For conductor 1, though, the energy required is $\frac{1}{2}c_{11}\phi_1^2$. Now raise the voltage on conductor 2 from zero to ϕ_2 . Additional work of $\frac{1}{2}c_{22}\phi_2^2$ is required. But an additional charge of $q_1 = c_{12}\phi_2$ also flows on to conductor 1. The work required to do this is $c_{12}\phi_1\phi_2$ (no integration is required since the potential is fixed at ϕ_1).

The total work is therefore the sum of the various terms

$$W = \frac{1}{2}c_{11}\phi_1^2 + \frac{1}{2}c_{22}\phi_2^2 + c_{12}\phi_1\phi_2 \quad (2.99)$$

But if we had reversed the order of charging the conductors, we would have arrived at the following result

$$W' = \frac{1}{2}c_{11}\phi_1^2 + \frac{1}{2}c_{22}\phi_2^2 + c_{21}\phi_2\phi_1 \quad (2.100)$$

But the energy of the system would surely be the same, or $W = W'$, which implies that $c_{12} = c_{21}$. Thus the capacitance matrix is symmetric. A symmetric matrix with non-zero

diagonal is invertible, thus justifying why we could go freely from the P matrix to its inverse C matrix.

Let us derive this result more directly as follows. The energy increment is given by $dW = dQv$, where we add charge dQ at constant voltage v . On conductor 1 we have

$$q_1 = c_{11}V_1 + C_{12}V_2 + \dots \quad (2.101)$$

$$dq_1 = c_{11}dV_1 + C_{12}dV_2 + \dots \quad (2.102)$$

So the energy increment is given by

$$dW = dq_1V_1 = c_{11}V_1dV_1 + C_{12}V_1dV_2 + \dots \quad (2.103)$$

Integration yields

$$W = \frac{1}{2}C_{11}V_1^2 + C_{12}V_1V_2 + \dots \quad (2.104)$$

Energy of the electrostatic field

Let us begin by finding the total energy for a distribution of point charges. We can imagine building up the distribution one charge at a time. The energy to bring in the first charge is naturally zero since the field is zero. The second charge, though, is repelled (or attracted) by the first charge so it requires more (less) energy to bring it in. In general the work required is given by

$$W_2 = Q_2\phi_{12} \quad (2.105)$$

where ϕ_{12} is the potential due to charge 1.

Since the electrostatic field is *conservative*, it does not matter how we bring in the second charge. Only its final position relative to the first charge is important. In terms of

$$W_2 = Q_2\phi_{12} = \frac{Q_2Q_1}{4\pi\epsilon R_{12}} \quad (2.106)$$

where R_{12} is the final distance between the point charges. Likewise, when we bring in the third charge, the extra work required is

$$W_3 = Q_3\phi_{13} + Q_3\phi_{23} = \frac{Q_3Q_1}{4\pi\epsilon R_{13}} + \frac{Q_3Q_2}{4\pi\epsilon R_{23}} \quad (2.107)$$

We can therefore write in general that the electrostatic energy takes on the following form

$$4\pi\epsilon W = \frac{Q_1Q_2}{R_{12}} + \frac{Q_1Q_3}{R_{13}} + \frac{Q_1Q_4}{R_{14}} + \dots + \frac{Q_2Q_3}{R_{23}} + \frac{Q_2Q_4}{R_{24}} + \dots \quad (2.108)$$

The general term has the form $\frac{Q_iQ_j}{4\pi R_{ij}}$, where i and j sum over all the particles in the system

$$W = \frac{1}{2} \sum \sum_{i \neq j} \frac{Q_iQ_j}{4\pi R_{ij}} \quad (2.109)$$

The factor of $\frac{1}{2}$ takes care of the double counting and enforcing $i \neq j$ ensures that we do not try to include the “self” energy of the particles.

We can rewrite the double sum into a more general form by observing that the inner sum is simply the potential due to all the particles evaluated at position of particle i

$$W = \frac{1}{2} \sum_{i \neq j} Q_i \sum_j \frac{Q_j}{4\pi R_{ij}} = \frac{1}{2} \sum Q_i \phi_i \quad (2.110)$$

If we now consider a charge distribution $\rho(r)$, it is easy to see how the above sum turns into an integral

$$W = \frac{1}{2} \int_V \rho \phi dV \quad (2.111)$$

The derived expression can be modified if we substitute $\rho = \nabla \cdot \mathbf{D}$ for charge

$$W = \frac{1}{2} \int_V \phi \nabla \cdot \mathbf{D} dV \quad (2.112)$$

and employ the chain rule

$$\nabla \cdot \phi \mathbf{D} = \phi \nabla \cdot \mathbf{D} + \mathbf{D} \cdot \nabla \phi \quad (2.113)$$

Since $\nabla \phi = -\mathbf{E}$, we have two volume integrals

$$W = \frac{1}{2} \int_V \nabla \cdot (\phi \mathbf{D}) dV + \frac{1}{2} \int_V \mathbf{D} \cdot \mathbf{E} dV \quad (2.114)$$

We can show that the first integral vanishes as follows. First apply the Divergence Theorem

$$\int_V \nabla \cdot (\phi \mathbf{D}) dV = \oint_S \phi \mathbf{D} \cdot \hat{\mathbf{n}} dS \quad (2.115)$$

Now take a surface S that is very large, say a large sphere. If the sphere is very large and the charge distribution is of finite extent, then at some great distance from the source the actual charge distribution is immaterial as only the net charge matters. We know that the radial potential and fields for a charge density take on the following limiting forms $\phi \sim \frac{1}{r}$ and $D \sim \frac{1}{r^2}$. Since the surface area $S \sim r^2$, the integrand vanishes.

Therefore the electrostatic energy is reduced to

$$W = \frac{1}{2} \int_V \mathbf{D} \cdot \mathbf{E} dV \quad (2.116)$$

If we define the energy density w , we have

$$W = \frac{1}{2} \int_V w dV \quad (2.117)$$

$$w = \frac{1}{2} \mathbf{D} \cdot \mathbf{E} \quad (2.118)$$

We found this to be true for an ideal capacitor but now we see this is true in general.

2.3 Non-linear capacitance

When capacitors are built from ideal conductors, the charge is always a linear function of the terminal voltages. But there are many situations when the charge on a body is a non-linear function of the terminal voltage, $Q = q(V)$. In this case we shall find that the derivative of charge with respect to voltage, dq/dV , plays an important role.

Small-signal capacitance

Here we consider a non-linear capacitor of the form $Q = q(v)$. Consider the differential charge dQ flowing onto a non-linear capacitor in response to a small-signal excitation. Let the terminal voltage consist of a DC large signal plus a small-signal AC voltage $V_s + v_s(t)$. A Taylor series expansion yields the incremental charge flow

$$Q(V) = q(V_s + v_s(t)) \approx q(V_s) + \left. \frac{dq(V)}{dV} \right|_{V=V_s} v_s(t) \quad (2.119)$$

The first term $q(V_s)$ is a constant, so if we break the charge into a static and dynamic component, we have

$$Q = Q_0 + \delta q \approx q(V_s) + \left. \frac{dq(V)}{dV} \right|_{V=V_s} v_s(t) \quad (2.120)$$

Since $Q_0 = q(V_s)$, we have

$$\delta q = \left. \frac{dq(V)}{dV} \right|_{V=V_s} v_s(t) = C v_s(t) \quad (2.121)$$

where we have defined the small-signal capacitance by

$$C \triangleq \left. \frac{dq(V)}{dV} \right|_{V=V_s} \quad (2.122)$$

Large-signal capacitance

For an arbitrary signal, we must continue to use the general expression $Q = q(v)$. In particular, to find the current we note that

$$i = \frac{dQ}{dt} = \frac{dq(v)}{dt} = \frac{dq}{dV} \frac{dV}{dt} = C(V) \frac{dV}{dt} \quad (2.123)$$

We see that the important term is again dq/dV . Assuming that $C(V)$ is analytic, we can perform a Taylor series expansion

$$i = (C_0 + C_1 V + C_2 V^2 + C_3 V^3 + \dots) \frac{dV}{dt} \quad (2.124)$$

Again, for small signals, the C_0 term will dominate, which leads again to the same definition of small-signal capacitance

$$C_0 = C(V_s) = \left. \frac{dq(V)}{dV} \right|_{V=V_s} \quad (2.125)$$

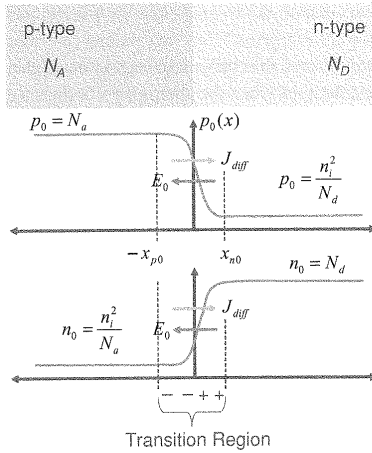


Figure 2.17 A pn-junction diode consisting of a n-type region with doping N_D and a p-type region with doping N_A . The concentration of electrons changes continuously from one majority carriers in the n-region to minority carriers in the p-region.

Junction diode

The capacitance of a pn-junction shown in Fig. 2.17 is a ubiquitous example of a non-linear capacitor. To calculate the capacitance of a reverse-biased junction diode, we need to find the charge density in a semiconducting material. As usual, assume that the p-type material is doped with N_A acceptor atoms and the n-type material is doped with N_D donor atoms. To start assume that no bias is applied across the terminals of the pn-junction.

From device physics, we know that when contact is made between these materials a built-in potential ϕ_{bi} builds up at the junction to prevent the diffusion of carriers across the junction. This potential develops in a small region around the junction, called the transition or depletion region. On the p-side of the depletion region, there are very few holes (hence the term *depletion*) and

$$\rho(x) \approx q(p_0 - N_A) < 0 \quad (2.126)$$

Note that the net charge is negative since the acceptors are negatively ionized. Likewise, on the n-side, the opposite is true

$$\rho(x) \approx q(-n_0 + N_D) > 0 \quad (2.127)$$

Since the distribution of carriers is approximately governed by Boltzmann statistics, we have

$$\rho(x) = \begin{cases} q(n_i e^{-q\phi_0(x)/kT} - N_A) & -x_{p0} < x < 0 \\ q(N_D - n_i e^{q\phi_0(x)/kT}) & 0 < x < x_{n0} \end{cases} \quad (2.128)$$

Poisson's equation arms us with the tools necessary to solve this problem since

$$\frac{d^2\phi}{dx^2} = -\frac{\rho(x)}{\epsilon_s} \quad (2.129)$$

This is a difficult problem to solve since the potential appears on both sides of the equation. There is a much simpler solution which yields excellent results. Let us simply assume that the transition region is fully depleted of free carriers of charge. Then we simply have

$$\rho(x) = \begin{cases} -qN_A & -x_{p0} < x < 0 \\ qN_D & 0 < x < x_{n0} \end{cases} \quad (2.130)$$

Now the problem is pretty easy. We can simply integrate twice, apply some boundary conditions, and we're done. Let's do the math

$$E_0(x) = -\frac{d\phi}{dx} = \int_{-x_{p0}}^x \frac{\rho(x')}{\epsilon_s} dx' + E_0(-x_{p0}) \quad (2.131)$$

The constant of integration is zero since the electric field outside of the transition region is zero. Thus

$$E_0(x) = -\frac{qN_A}{\epsilon_s}(x + x_{p0}) \quad (2.132)$$

The electric field drops linearly as we traverse the junction. This is in contrast to a linear capacitor with constant electric field. This is totally anticipated by Gauss' Law without any recourse to mathematics. Gauss' Law tells us that electric fields diverge on charge. Therefore a uniform charge density will eat away at the field in a linear manner!

We can also do the integral from the n-side of the junction

$$E_0(x_{n0}) = 0 = \int_x^{x_{n0}} \frac{\rho(x')}{\epsilon_s} dx' + E_0(x) = \frac{qN_D}{\epsilon_s}(x_{n0} - x) + E_0(x) \quad (2.133)$$

or

$$E_0(x) = \frac{-qN_D}{\epsilon_s}(x_{n0} - x) \quad (2.134)$$

Since the field is continuous, at $x = 0$ we have

$$qN_A x_{p0} = qN_D x_{n0} \quad (2.135)$$

which simply states that the total charge on the n-side equals the total charge on the p-side, which is somewhat obvious. Gauss' Law guarantees this fact, since the electric field is zero outside the transition region. A plot of the electric field in the pn-junction is shown in Fig. 2.18.

If we now integrate again to arrive at potential, we find that on the p-side

$$\phi_0^p(x) = \phi_p + \frac{qN_A}{2\epsilon_s}(x + x_{p0})^2 \quad (2.136)$$

and on the n-side

$$\phi_0^n(x) = \phi_n - \frac{qN_D}{2\epsilon_s}(x - x_{n0})^2 \quad (2.137)$$

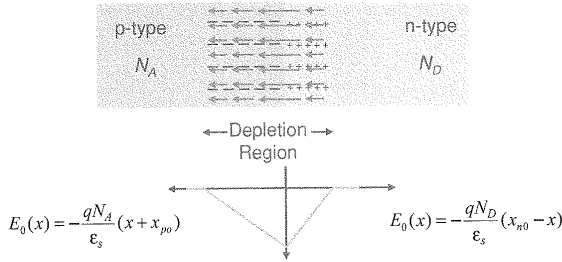


Figure 2.18 The built-in electric fields in a pn-junction diode.

At the interface, the potential must be a continuous function. This leads to

$$\phi_n - \frac{qN_D}{2\epsilon_s}x_{n0}^2 = \phi_p + \frac{qN_A}{2\epsilon_s}x_{p0}^2 \quad (2.138)$$

The above equation, along with Eq. (2.135), allows us to solve for the depletion region depths

$$x_{n0} = \sqrt{\frac{2\epsilon_s\phi_{bi}}{qN_D} \left(\frac{N_A}{N_A + N_D} \right)} \quad (2.139)$$

$$x_{p0} = \sqrt{\frac{2\epsilon_s\phi_{bi}}{qN_A} \left(\frac{N_D}{N_A + N_D} \right)} \quad (2.140)$$

where $\phi_{bi} = \phi_n - \phi_p > 0$.

Now let's apply a reverse bias, or a potential that tends to increase the barrier height between the n- and p-regions. Certainly the current flow is essentially zero. In fact, for zero reverse bias, the built-in potential is precisely large enough to prevent majority carriers in one region from diffusing into the other region. But the barrier cannot prevent minority carriers from crossing the junction. So under a reverse bias, minority carriers that are thermally generated within a few diffusion constants from the junction will in fact give rise to a reverse current. This reverse current is essentially independent of the applied bias (until breakdown occurs). To see this intuitively, imagine blind mole rats that live near a cliff. Every once in a while a mole rat will come out of his hole and run along in a random direction. If he is so unlucky to run towards the cliff, then surely he will plunge to his death. Now, if we make the cliff twice as tall, the rate at which mole rats die will not change!

So back to the reverse biased pn-junction. When we apply an additional reverse bias, this voltage must be dropped across the transition region. This can only happen if the transition region *grows* in extent. In fact, we can now see that the equations derived under zero bias can be reused if we simply replace ϕ_{bi} with $\phi_{bi} - V_D$, where V_D is the additional reverse bias applied to the pn-junction. The depletion region can be written as a function of the

reverse bias

$$x_n(V_D) = \sqrt{\frac{2\epsilon_s(\phi_{bi} - V_D)}{qN_D} \left(\frac{N_A}{N_A + N_D} \right)} = x_{n0} \sqrt{1 - \frac{V_D}{\phi_{bi}}} \quad (2.141)$$

$$x_p(V_D) = \sqrt{\frac{2\epsilon_s(\phi_{bi} - V_D)}{qN_A} \left(\frac{N_D}{N_A + N_D} \right)} = x_{p0} \sqrt{1 - \frac{V_D}{\phi_{bi}}} \quad (2.142)$$

The total depletion region depth is given by

$$X_d(V_D) = x_p(V_D) + x_n(V_D) = \sqrt{\frac{2\epsilon_s(\phi_{bi} - V_D)}{q} \left(\frac{1}{N_D} + \frac{1}{N_A} \right)} = X_{d0} \sqrt{1 - \frac{V_D}{\phi_{bi}}} \quad (2.143)$$

As we increase the reverse bias voltage, the depletion region grows to accommodate more charge. The charge on either side of the junction is given by

$$Q_j(V_D) = -qN_Ax_p(V_D) = -qN_AX_{d0} \sqrt{1 - \frac{V_D}{\phi_{bi}}} \quad (2.144)$$

Charge is not a linear function of voltage, and so we have a non-linear capacitor! To see that this is indeed a capacitor, note that the pn-junction is initially neutral. When we apply a reverse bias voltage, the depletion region grows, exposing ionized donor and acceptor ions. Since the depletion region is free of mobile charge, a corresponding charge must be transferred from the battery in order to precisely balance the exposed charge. Hence, from the terminals of the battery, it appears that we have connected a non-linear capacitor.

What is the small-signal capacitance? Differentiating the expression for charge we arrive at

$$C_j = \frac{dQ_j}{dV} \bigg|_{V=V_D} = \frac{d}{dV} \left(-qN_Ax_{p0} \sqrt{1 - \frac{V}{\phi_{bi}}} \right) \bigg|_{V=V_D} \quad (2.145)$$

or

$$C_j = \frac{C_{j0}}{\sqrt{1 - \frac{V_D}{\phi_{bi}}}} \quad (2.146)$$

where we have defined

$$C_{j0} = \frac{qN_Ax_{p0}}{2\phi_{bi}} = \frac{qN_A}{2\phi_{bi}} \sqrt{\left(\frac{2\epsilon_s\phi_{bi}}{qN_A} \right) \left(\frac{N_D}{N_A + N_D} \right)} = \sqrt{\frac{2\epsilon_s}{2\phi_{bi}} \frac{N_A N_D}{N_A + N_D}} \quad (2.147)$$

Don't bother to memorize the expression for C_{j0} . You can derive it very easily if we notice that

$$C_{j0} = \epsilon_s \sqrt{\frac{q}{2\epsilon_s\phi_{bi}} \left(\frac{1}{N_D} + \frac{1}{N_A} \right)^{-1}} = \frac{\epsilon_s}{X_{d0}} \quad (2.148)$$

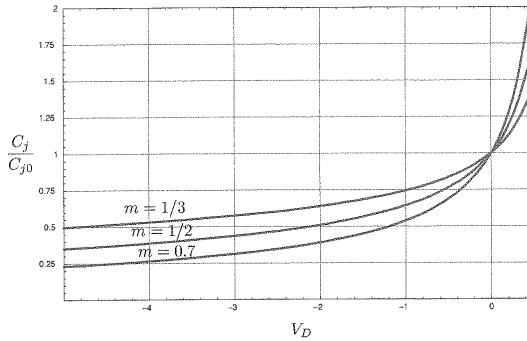


Figure 2.19 The normalized diode junction capacitance as a function of the terminal voltage for three different doping profiles. The $m = 1/2$ corresponds to the abrupt junction.

which is surprisingly the formula for a parallel plate capacitor! So the junction capacitance is simply

$$C_j(V_D) = \frac{\epsilon_s}{X_d(V_D)} \quad (2.149)$$

Again we can explain the simplicity of this result. When we apply an incremental voltage, the increment of charge comes from the outer edges of the depletion region, since this is the direction of the growth of the depletion region. Thus it appears that we have a parallel plate capacitor with plate separation of $X_d(V)$.

This variation of this capacitance with voltage is plotted in Fig. 2.19. The shape of the curve can be altered by proper adjustment of the junction doping profile. In the above analysis we made the assumption that the doping profile is a step function across the junction. In practice, we assume that $X_d = X_{d0} \left(1 - \frac{V_D}{\phi_{bi}}\right)^m$ and experimentally determine the “grading coefficient” m .

MOS capacitor

The MOS capacitor, or MOS-C structure, shown in Fig. 2.20, is a capacitor constructed with one metal gate, an oxide insulator, and another semiconducting “plate.” In practice, we often employ a poly-silicon gate, which acts like a metal. Let us assume that we short the terminals of the capacitor. Here the body is p-type, whereas the gate is a metal or heavily doped n-type material. Upon contact, electrons leave the gate and flow into the p-type material and the charge is negative. Likewise, holes from the body come into the gate and the charge is positive. By allowing the materials to exchange charge, we can see that a built-in potential difference will arise between the materials until an equilibrium potential difference has been established.

Since all systems strive to reside in a minimum potential state, it is not surprising to find the positive charges on the gate residing above a depletion formed in the body, as shown in Fig. 2.21a. Remember that this is the equilibrium state. Initially free charge in the body near the contact will flow into the gate, forming a transient depletion region near the wire

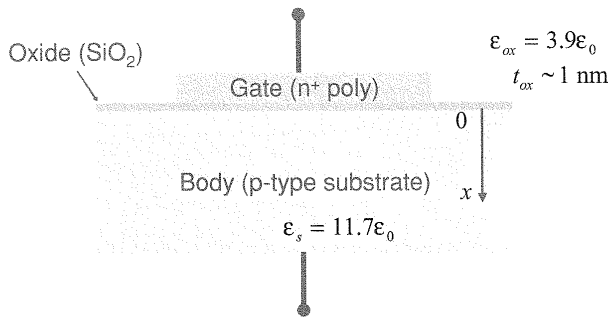


Figure 2.20 The cross section of a MOS-C capacitor structure. Permeability numbers are for Si and SiO₂.

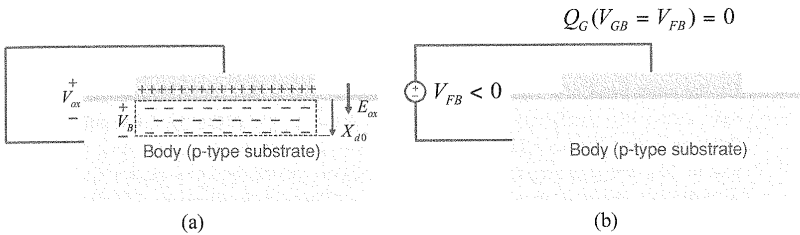


Figure 2.21 (a) The MOS-C structure under a zero-bias voltage. (b) The MOS-C under a flat-band voltage bias.

contact. But in the steady state, other holes will occupy these states, leaving the higher energy region under the gate depleted of carriers.

Flat band

Now if we apply a negative voltage V_{FB} to cancel the internal built-in potential, as shown in Fig. 2.21b, we see that the charge on the gate and the body will go to zero. This is the so-called “flat-band” condition, since the electric field (and hence the band bending) along the MOS interface is zero. The flat-band voltage necessary to do this is simply

$$V_{FB} = -(\phi_{n^+} - \phi_p) \quad (2.150)$$

for the p-type body and n^+ gate material.

Accumulation

If we further decrease the potential beyond the “flat-band” condition, we essentially have a parallel plate capacitor. As shown in Fig. 2.22a, since a negative bias attracts holes from the body to flow under the gate, and since holes are majority carriers, plenty of holes are

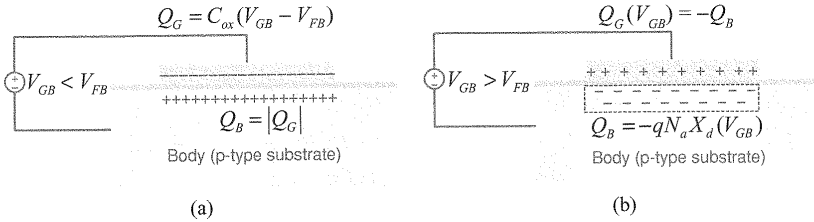


Figure 2.22 (a) The MOS-C structure in the accumulation region. (b) The MOS-C in the depletion region.

available to do this. Likewise plenty of electrons can flow into the gate through the n^+ material. Operated in a region with bias $V_{GB} < V_{FB}$ is called “accumulation,” since holes accumulate at the surface of the MOS-C structure.

Depletion

On the other hand, if we apply a positive voltage $V_{GB} > V_{FB}$, as shown in Fig. 2.22b, we have a case very similar to the equilibrium condition. The potential at the gate is higher than the body. The body charge consists of mostly ionized carriers forming a depletion region. The potential drop is across the oxide region and the depletion region. In this region of operation, the number of minority carriers in the body is simply too small to have an appreciable impact on the behavior of the device.

Inversion

As we continue to apply a larger and larger gate voltage, we can see that the *surface potential* will continue to increase. Since the electron concentration is proportional to the potential, at the surface we have

$$n_s = n_i e^{\frac{q\phi_s}{kT}} \quad (2.151)$$

If the surface potential becomes large enough, the minority carrier concentration can become large enough to “invert” the material doping, as shown in Fig. 2.23a. When this occurs, the depletion region essentially stops growing and the extra charge is now readily provided by the inversion charge at the surface. For this reason, the potential at the surface is also “pinned” since electrons flow to the surface of the MOS-C structure and any extra voltage drop occurs in the oxide region. The gate voltage required to produce this surface inversion is called the threshold voltage V_T .

Threshold voltage

The threshold voltage for the p-body MOS-C structure is defined as the gate–body voltage necessary to cause the surface of the body to change from p-type to n-type. For this condition

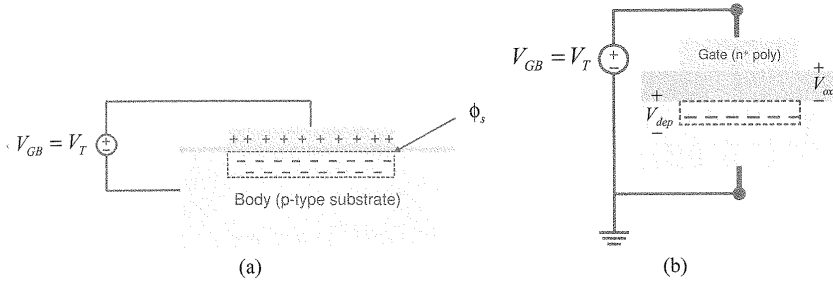


Figure 2.23 (a) The MOS-C structure in the inversion region. (b) The gate–body voltage consists of the built-in voltage, the oxide voltage, and the voltage drop across the depletion region.

to occur, the surface potential has to equal the negative of the p-type potential. Let's apply KVL around the loop shown in Fig. 2.23b

$$V_{GS} = V_{FB} + V_{ox} + V_{dep} \quad (2.152)$$

Note that $\phi_s = V_{BS} = -2\phi_p$ to cause inversion. The factor of two is somewhat confusing since $n_s = N_A$ implies

$$n_s = n_i e^{\frac{q\phi_s}{kT}} = N_A \quad (2.153)$$

or $\phi_s = -\phi_p$. So the potential difference between the inverted surface and the body is $-\phi_p - \phi_p = -2\phi_p$. Since the body is grounded, this is precisely the potential we need at the surface to invert it.

The term V_{ox} can be replaced by applying the normal field boundary condition

$$V_{ox} = E_{ox} t_{ox} = \frac{\epsilon_s}{\epsilon_{ox}} t_{ox} E_s \quad (2.154)$$

The electric field at the surface of the silicon can be found by application of Gauss' Law to the depletion region. Since we assume that the electric field is zero inside the body, the electric flux at the surface must equal the charge stored inside the depletion region

$$E_s = \frac{q N_A x_{dep}}{\epsilon_s} = \frac{q N_A}{\epsilon_s} \sqrt{\frac{2\epsilon_s}{q N_A} \phi_s} = \sqrt{\frac{2q N_A (-2\phi_p)}{\epsilon_s}} \quad (2.155)$$

These substitutions yield the following value for the threshold voltage

$$V_{TH} = V_{FB} - 2\phi_p + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_A (-2\phi_p)} \quad (2.156)$$

MOS Q - V and C - V curves

Given the above calculations, we can derive the MOS Q - V curve. A simple approximation is to assume that once inversion occurs, the depletion region stops growing. This is a good approximation since charge is an exponential function of the surface potential. Under this

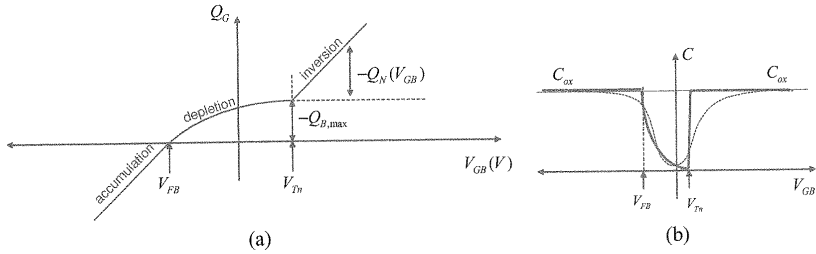


Figure 2.24 (a) The MOS-C charge voltage or Q - V curve. (b) The MOS-C capacitance versus voltage curve. Note this is simply a derivative of the Q - V curve.

assumption

$$Q_G(V_{TH}) \approx -Q_{B,max} \quad (2.157)$$

or

$$Q_G(V_{GB}) = C_{ox}(V_{GB} - V_{TH}) - Q_{B,max} \quad (2.158)$$

The MOS Q - V curve is shown in Fig. 2.24a. In accumulation $V_{GB} < V_{FB}$, we have neglected the minority carriers and thus the charge is proportional to the applied bias, much like a regular linear capacitor. But note carefully that the charge is not zero for $V_{GB} = 0$ but rather for V_{FB} . In the region between flat band and threshold, $V_{FB} < V_{GB} < V_{TH}$, we are in the depletion region. Again, if we ignore the effects of the minority carriers, the charge is due to the depletion region only. Then at $V > V_{TH}$ we assume that the depletion region grows imperceptibly and all of the charge appears at the surface of the body through the inversion charge.

The MOS-C small-signal C - V curve is derived by simply taking the derivative of the Q - V curve. This is shown in Fig. 2.24b. In the accumulation and inversion regions, the capacitance is essentially given by C_{ox} , or the parallel plate capacitance of the MOS-C. In the depletion region, the capacitance drops from C_{ox} as the depletion region grows larger and larger. But then suddenly, and quite artificially, inversion occurs and the capacitance jumps back to the value of C_{ox} . Of course, in reality, the minority carriers do not play a negligible role in the MOS-C structure. Below but near the threshold voltage, their presence is very important for the small-signal capacitance, even if their numbers are small, since the rate at which these charges appear with increasing voltage is exponential. A more realistic C - V curve, as sketched with the dashed line in the figure, shows a smooth transition across inversion and flat band.

In accumulation mode the capacitance is just due to the voltage drop across t_{ox} . In inversion, the incremental charge comes from the inversion layer (the depletion region stops growing). In the depletion region, the voltage drop is across the oxide and the depletion region, so we have two series connected capacitors. To calculate the capacitance in the depletion region directly, note that an equivalent circuit for the MOS-C structure in this region is

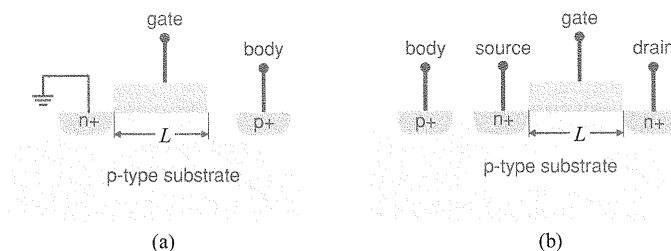


Figure 2.25 (a) An integrated MOS-C structure is constructed with a surface body contact and an n^+ “source” layer to provide electrons for inversion. (b) A MOSFET is a natural inversion mode MOS-C structure. If the source and drain are tied together, both regions can become a source for electrons.

two series capacitors, an oxide capacitance C_{ox} in series with the depletion capacitance

$$C_{dep} = \frac{\epsilon_s}{x_{dep}} \quad (2.159)$$

Taking the capacitors in series, we have

$$C_{tot} = \frac{C_{dep}C_{ox}}{C_{dep} + C_{ox}} = \frac{C_{ox}}{1 + \frac{\epsilon_s}{\epsilon_{ox}} \frac{t_{ox}}{x_{dep}}} \quad (2.160)$$

In the above discussion you may have wondered, “Where do the electrons come from to form the inversion layer?” In the body of the MOS-C structure, electrons are minority carriers and few and far between. So when inversion occurs, where do we find all the electrons necessary to invert the surface? Well, there was a subtle assumption that if we apply a change in gate voltage, we wait long enough for thermal generation to create a sufficient number of electrons to form the surface layer. We may have to wait a very long time! In other words, if we apply a fast enough signal to the gate, there isn’t enough time for the minority carriers to be generated and thus the capacitance remains at the low value given by depletion.

While the depletion region can respond very quickly to our gate voltage since it is formed by majority carriers, the minority carrier generation is slow. There is a simple way to solve this problem, as shown in Fig. 2.25, where a n^+ grounded contact is placed adjacent to the gate. Normally electrons are prevented from entering the body, like any good pn-junction. But as we raise the surface potential, electrons can easily diffuse into the surface of the structure. Since the energy distribution of electrons in thermal equilibrium is exponential,² changing the potential barrier linearly results in an exponential increase in the number of electrons that can cross the n^+ -surface junction and likewise an exponential increase in surface conductors.

There are a few physical effects that we have neglected to mention. The most important physical effects are the poly-silicon depletion region and the quantum mechanical effects. The poly-depletion region is easy enough to understand. Since the gate material is usually

² Or, if you like, the tail of the distribution is approximately exponential.

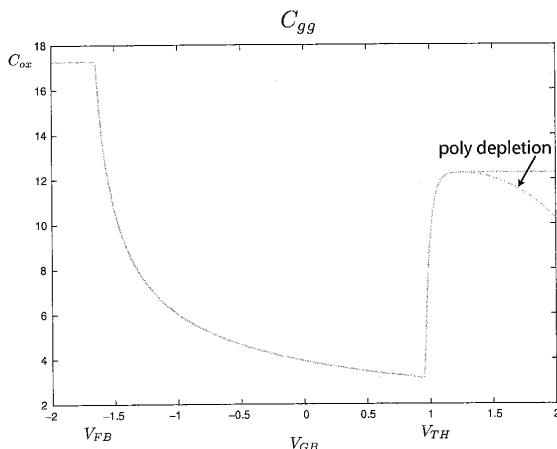


Figure 2.26 A more realistic MOS C - V curve takes into account the role of minority carriers below threshold, the quantum mechanical confinement, and the poly-silicon gate depletion.

constructed with a heavily doped poly-silicon, rather than a metal, it will also form a small depletion region. As the device t_{ox} is reduced in short-channel devices, this depletion region width is an appreciable fraction of t_{ox} , resulting in a lower effective gate bias. As shown in Fig. 2.26, as we increase the V_{GB} beyond V_T , the depletion region width tends to increase, lowering the capacitance.

The quantum mechanical confinement is also easily understood if we solve Poisson's equation along with Schrodinger's equation in a consistent fashion. The result is that the peak charge does not occur at the surface of the silicon body, but slightly displaced from the surface. In effect, the oxide thickness increases, also resulting in a smaller capacitance.

2.4 References

Some of the material was put together while teaching an Electromagnetics course at UC Berkeley. Useful references include Feynman's Lectures [18], Cheng [6], Reitz [53], and Inan and Inan's book [62]. Some of the device physics material was drawn from lecture notes put together by Howe and Sodini [49].

3 Resistance

You may be tempted to skip this chapter. After all, resistors are pretty elementary, right? Actually, below the surface resistors are complicated beasts. In fact, resistors apparently defy one of the Newton's fundamental law of physics, as charges in resistors move with constant velocity when subjected to an external force. We will probe into the inner workings of resistors in order to answer this riddle.

3.1 Ohm's Law

Let's begin with the venerable Ohm's Law. Is it trivial that $V = I \times R$? Perhaps all that is going on is that

$$V = f(I) = f(0) + f'(0)I + \frac{1}{2}f''(0)I^2 + \dots \approx f'(0)I \quad (3.1)$$

where we have kept the linear term in the Taylor series expansion of the voltage-current relationship.¹ If this is the case, what's the range of validity of the above approximation. What's surprising is that Ohm's Law is valid over an enormous range.

Note that Ohm's Law is also equivalently expressed as

$$\mathbf{J} = \sigma \mathbf{E} \quad (3.2)$$

We assert that the ratio between the voltage and current in a conductor is always constant. We see this stems from $J = \sigma E$

$$I = \int_A \mathbf{J} \cdot d\mathbf{S} = \int_A \sigma \mathbf{E} \cdot d\mathbf{S} \quad (3.3)$$

$$V = \int_C \mathbf{E} \cdot d\mathbf{l} \quad (3.4)$$

The ratio is called the resistance

$$R = \frac{V}{I} = \frac{\int_C \mathbf{E} \cdot d\mathbf{l}}{\int_A \sigma \mathbf{E} \cdot d\mathbf{S}} \quad (3.5)$$

¹ The expansion point is about zero current since we experimentally observe that the average current in a conductor is zero for zero applied voltage.

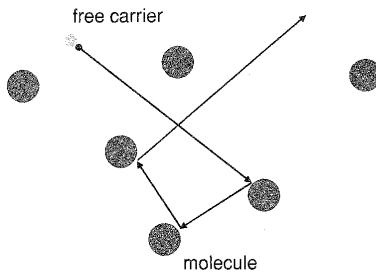


Figure 3.1 Simple billiard-ball model of collision between a free carrier and a molecule.

What is disturbing about Ohm's Law is that current, which is proportional to the velocity, is proportional to the electric field, which is imparting a force on the charge. Is Newton rolling in his grave?

Conductivity of a gas

Electrical conduction is due to the motion of positive and negative charges. For water with $\text{pH} = 7$, the concentration of hydrogen H^+ ions (and OH^-) is

$$10^{-7} \text{ mole/L} = 10^{-10} \text{ mole/cm}^3 = 10^{-10} \times 6.02 \times 10^{23} \text{ cm}^{-3} = 6 \times 10^{13} \text{ cm}^{-3} \quad (3.6)$$

Typically, the concentration of charged carriers is much smaller than the concentration of neutral molecules. The motion of the charged carriers (electrons, ions, molecules) gives rise to electrical conduction.

At a temperature T , each charged carrier will move in a random direction and velocity until it encounters a neutral molecule or another charged carrier. Since the concentration of charged carriers is much less than molecules, it will most likely encounter a molecule. For a gas, the molecules are widely separated (~ 10 molecular diameters). After colliding with the molecule, there is some energy exchange and the charge carrier will come out with a new velocity and new direction. For simplicity, assume elastic collisions and ignore atomic level force interactions.

This is shown schematically in Fig. 3.1. The key observation is that the initial velocity and direction is lost (randomized) after a few collisions. When we apply an electric field, during each "free flight," the carriers will gain a momentum of $q\mathbf{E} \times t$. Therefore, after t seconds, the momentum is given by

$$M\mathbf{u} + q\mathbf{E}t \quad (3.7)$$

If we take the average momentum of all particles at any given time, we have

$$M\bar{\mathbf{u}} = \frac{1}{N} \sum_j (M\mathbf{u}_j + q\mathbf{E}t_j) \quad (3.8)$$

where we assume N carriers, an initial momentum $M\mathbf{u}_j$ before the collision, and a time t_j between collisions.

When we sum over all the initial random velocities of the particles, we are averaging over a large number of random variables with zero mean, and so the average is zero

$$M\bar{\mathbf{u}} = \frac{1}{N} \sum_j \left(M\mathbf{u}_j + \mathbf{E}qt_j \right) \quad (3.9)$$

The momentum gain is therefore given by the second term

$$M\bar{\mathbf{u}} = \frac{1}{N} \sum_j \mathbf{E}qt_j = \mathbf{E}q\tau \quad (3.10)$$

where τ is the average time between collisions.

Current is defined as the amount of charge crossing a particular cross-sectional area per unit time. Consider a density of N carriers carrying a charge q and moving with velocity v_d . How many cross a perpendicular surface A ? In a time Δt , carriers move a distance of $v_d\Delta t$. So any carriers in the volume $Av_d\Delta t$ will cross the surface if they are all moving towards it. The current is therefore

$$I = \frac{Av_d\Delta t Nq}{\Delta t} = v_d NqA \quad (3.11)$$

The vector current density $\mathbf{J}$ is therefore

$$\mathbf{J} = Nq\mathbf{v}_d \quad (3.12)$$

Using Eq. (3.10) we arrive at

$$\mathbf{J} = Nq\bar{\mathbf{u}} = Nq \left(\frac{\mathbf{E}q\tau}{M} \right) = Nq^2 \frac{\tau}{M} \mathbf{E} = \sigma \mathbf{E} \quad (3.13)$$

Since current is contributed by positive and negative charge carriers

$$\mathbf{J} = \mathbf{J}^+ - \mathbf{J}^- = e \left(\frac{N^+ e \tau^+}{M^+} - \frac{-N^- e \tau^-}{M^-} \right) \mathbf{E} \quad (3.14)$$

or

$$\mathbf{J} = \sigma \mathbf{E} \quad (3.15)$$

The conductivity of a gas is therefore given by

$$\sigma = e^2 \left(\frac{N^+ \tau^+}{M^+} + \frac{N^- \tau^-}{M^-} \right) \quad (3.16)$$

Conduction in metals

The high conductivity of metals is due to a large concentration of free electrons. These electrons are not attached to the solid but are free to move about the solid. This sea of electrons, or an electron gas, in effect keeps the solid together. In metal sodium, for instance, each atom contributes a free electron. The ionized molecules are evenly spaced and form a

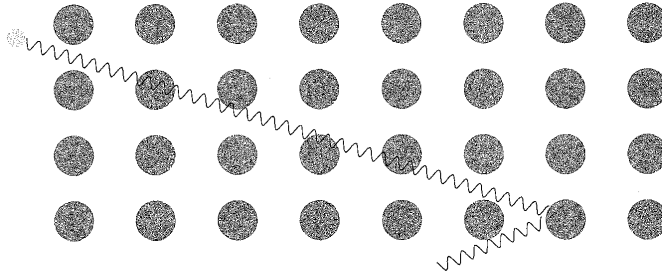


Figure 3.2 An electron can penetrate through potential barriers of ions due to the periodic nature of the crystal.

crystal lattice. From the measured value of conductivity (easy to do), we can back calculate the mean free time

$$\tau = \frac{\sigma m}{Ne^2} = \frac{(1.9 \times 10^7)(9 \times 10^{-28})}{(2.5 \times 10^{22})(23 \times 10^{-20})} = 3 \times 10^{-12} \text{sec} \quad (3.17)$$

This value of mean free time is surprisingly long. The mean velocity for an electron at room temperature is roughly given by

$$\frac{mv^2}{2} = \frac{3}{2}kT \quad (3.18)$$

or $v = 3 \times 10^7 \text{cm/sec}$. At this speed, the electron travels a distance $v\tau = 3 \times 10^{-7} \text{cm}$. The molecular spacing between adjacent ions is only about $3.8 \times 10^{-8} \text{cm}$. Why is it that the electron is on average zooming by ten positively charged ions?

The answer to this deep puzzle requires us to admit the wave nature of electrons. As shown schematically, in Fig. 3.2, the free carrier can penetrate right through positively charged host atoms. Quantum mechanics explains all of this. For a periodic arrangement of potential fluctuations, the electron does not scatter. It is as if our electron is free to travel around the crystal, but its mass is not the same as the electron mass, but an “effective” mass determined by the crystal. So why does the electron scatter at all?

First, at a temperature T , the atoms are in random motion and thus upset the periodicity, as shown in Fig. 3.3a. Even at low temperatures, though, the presence of impurities upsets the periodicity, as seen in Fig. 3.3b.

Free carrier mobility

Drift

As we have seen, when a carrier in a conductor is accelerated by an external field, it initially travels in the direction of the field. But after scattering with an impurity or phonon, the electron's new trajectory is random and independent of the field. We have shown that the average velocity of carriers increases proportional to the field E

$$v_d = \frac{q\tau_c}{m} E = \mu E \quad (3.19)$$

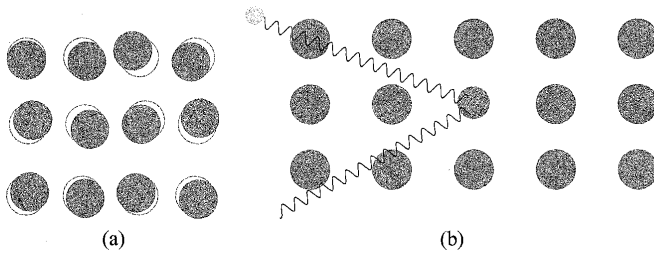


Figure 3.3 Electron scattering occurs due to (a) the random vibration of the molecules which upset the periodicity of the crystal lattice and (b) due to the presence of impurities in the lattice.

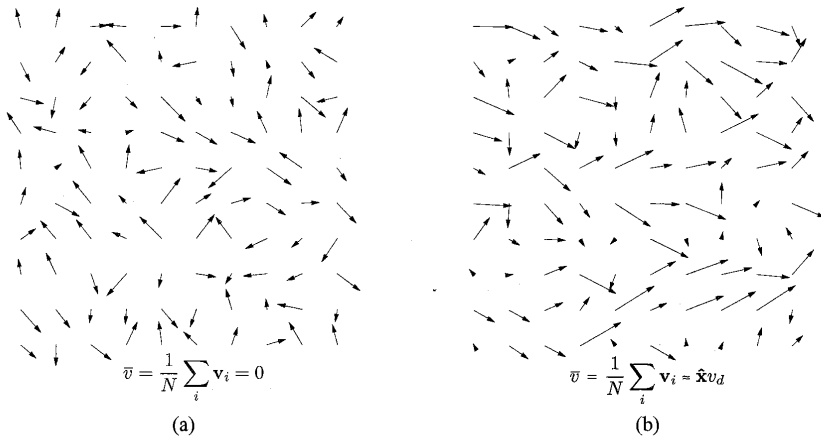


Figure 3.4 (a) In a conductor in the absence of any external fields, free carriers are moving with random velocity in random directions. (b) In response to an external electric field, there is a net motion, or drift, in the direction of the field.

The constant μ is the mobility, and the “drift” velocity v_d is so named because most of the motion is still random as $v_d \ll v_{th}$.

In Fig. 3.4a, in the absence of an external field, electrons are moving with random velocity (net motion is zero). In Fig. 3.4b, electrons are moving with random velocity plus a drift component to the right (drift has been exaggerated).

Charge conservation

One of the fundamental facts of nature is charge conservation. This has been experimentally observed and verified in countless experiments. Therefore, if we look at a volume bounded by surface S , as shown in Fig. 3.5, if there is net current leaving this volume, it must be

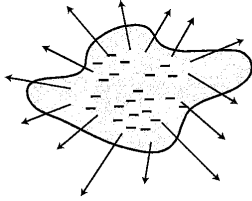


Figure 3.5 By charge conservation, in a region with net positive current flow *out* of the region, a net negative charge builds up inside over time.

accompanied by a charge pile-up in the region of opposite polarity

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = -\frac{\partial}{\partial t} \int_V \rho dV \quad (3.20)$$

We can transform this simple observation by employing the Divergence Theorem

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{J} dV \quad (3.21)$$

For any volume we have found that

$$\int_V \nabla \cdot \mathbf{J} dV - \frac{\partial}{\partial t} \int_V \rho dV = \int_V \left(\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} \right) dV = 0 \quad (3.22)$$

If this is true for any volume V , it must be true that

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \quad (3.23)$$

Since $\mathbf{J} = \sigma \mathbf{E}$, we have

$$\nabla \cdot \sigma \mathbf{E} + \frac{\partial \rho}{\partial t} = 0 \quad (3.24)$$

If the conductivity is constant, we have

$$\sigma \nabla \cdot \mathbf{E} + \frac{\partial \rho}{\partial t} = 0 \quad (3.25)$$

It is experimentally observed that Gauss' Law is satisfied in all reference frames of constant velocity so $\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon}$

$$\frac{\sigma}{\epsilon} \rho + \frac{\partial \rho}{\partial t} = 0 \quad (3.26)$$

If we assume that ρ is a function of time only, we have the following differential equation

$$\frac{\sigma}{\epsilon} \rho + \frac{d\rho}{dt} = 0 \quad (3.27)$$

Letting $\tau = \sigma/\epsilon$, we solve the above equation

$$\rho = \rho_0 e^{-t/\tau} \quad (3.28)$$

Thus the charge density in a conductor decays exponentially to zero with a rate determined by the time constant $\tau = \sigma/\epsilon$, or the *relaxation time* of the material. Earlier we deduced that conductors have zero volume charge density under electrostatic conditions. Now we can estimate the time it takes to reach this equilibrium.

Relaxation time for good conductors

In fact, a good conductor like copper has $\sigma \sim 10^8 \text{ S/m}$. It is hard to measure the dielectric constant of a good conductor, since the effect is masked by the conductive response. For $\epsilon \sim \epsilon_0$, we can estimate the relaxation time

$$\tau \sim \frac{10 \times 10^{12}}{10^8} \text{ s} \sim 10^{-19} \text{ s} \quad (3.29)$$

This is an *extremely* short period of time! How do electrons move so fast? Note this result is independent of the size of the conductor. The answer is that they don't. Only a slight displacement in the charge distribution can balance the net charge and thus transfer it to the border region.

3.2 Conduction in semiconductors

Semiconductors have formed the foundation for modern society. Historians may call our age the “sand age” since virtually all of our machines are made of Si semiconductors. The conductivity of semiconductors, as their name suggests, lies somewhere between conductors and insulators. We will explain this with a very simple band model.

From quantum mechanics we know that the allowed energy levels for an atom are discrete (two electrons can occupy a state with opposite spin). When atoms are brought into close proximity, these energy levels split. If there are a large number of atoms, such as in a solid crystal, the discrete energy levels form a seemingly continuous band. In our simple model, each material solid is characterized by a valence band and a conduction band, with a forbidden region called the band gap.

The valence band corresponds to bound electronic states. In other words, electrons with this energy level are bound to their host atoms. Usually these electrons are part of a covalent bond. Conduction band electrons, on the other hand, are free to move around in the solid. Of course they are not free to leave the material as there is an additional amount of energy required for an electron to leave the solid. This is characterized by the work function of the material. The energy distribution of the electrons is determined by Fermi-Dirac statistics and the Fermi level E_F of the material. At low temperature, most electrons fill energy levels below the Fermi level.

One of the most important properties of a solid is the energy difference between the valence band and the conduction band, or the band gap. This is important because a valence band electron can become a conduction band electron if it can attain this energy level, either through incoming radiation (a photon), or through thermal energy. If the band gap is very small, then even at low temperatures, many electrons will occupy the conduction band.

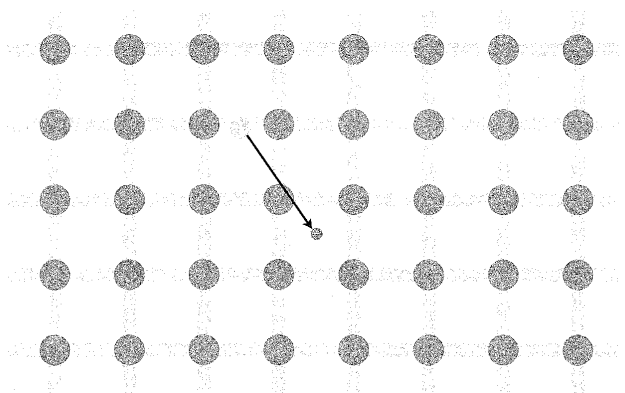


Figure 3.6 A cartoon two-dimensional Si lattice. Circles represent Si atoms and lines represent covalent bonds formed by valence band electrons. Broken bonds give rise to electrons and holes.

Insulators have a very large band gap, about 3–7 eV, so that the probability that an electron has sufficient energy to occupy the conduction band is extremely small. And semiconductors have an intermediate band gap. Si, for instance, has a band gap of about 1.12 eV. This is actually quite large compared to the thermal energy of kT , which is about 26meV at room temperature.

Electrons and holes

When early researchers measured the conductivity of Si and other semiconductors, they were in for a surprise. The conductivity numbers were often a factor of 2 from their predictions. It was as if there were twice as many free carriers in the material. Well, in fact, that is precisely what happens in a semiconductor. In addition to electrons, “holes” also conduct current.

In Fig. 3.6, we have a cartoon diagram of a two-dimensional crystal such as Si that forms four covalent bonds with its neighbors, represented by lines in the figure. In this figure valence band electrons are the covalent bonds in the crystal. If we break one of these bonds to free an electron, there is in essence a hole of positive charge left behind. It is important to realize that this hole is in fact free to move. This is because it represents a state of low energy for another electron. So in fact it is very easy for another nearby valence band electron to hop into this hole. In doing so it creates another hole and thus the hole travels! So electrons moving in the valence band from bond to bond corresponds to the equivalent motion of a positive charge. A very crude analogy is a bubble in a sealed liquid. This bubble corresponds to our hole, and it moves around the liquid as if it were a real particle.

When a conduction band electron encounters a hole, the process is called recombination. The electron and hole annihilate one another thus depleting the supply of carriers. In thermal equilibrium, a generation process counterbalances to produce a steady stream of carriers. We can see that the thermal generation of free carriers produces an equal amount of electrons

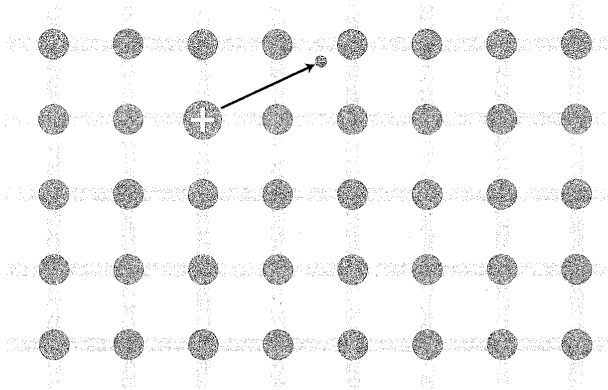


Figure 3.7 A donor dopant atom in the Si lattice contributes a free electron.

and holes, or $n_0 = p_0$. If the thermal generation rate G_{th} is equal to the recombination rate R , we have

$$G_{th} = R = kn \times p \quad (3.30)$$

where we assert that the regeneration rate is proportional to the product of holes and electrons. This is logical because for every electron the probability of recombination is proportional to the number of holes, and vice versa. Since at a given temperature the thermal generation rate is constant, the product of $n \times p$ is also a constant

$$np = n_i^2 \quad (3.31)$$

which is the intrinsic concentration of electrons (and holes) in a material at a given temperature. For Si at 300K, this is about 10^{10}cm^{-3} .

Doping

Thus far our discussion of semiconductors, in particular Si, has been limited to pure or intrinsic Si. In practice, pure Si is very hard to produce as even trace quantities of impurities can dramatically alter the behavior of the resulting material. This is in fact one of the great benefits of Si in that its behavior is very well controlled through selective doping.

Imagine adding a trace amount of a group V element such as phosphorous or arsenic to a Si crystal, as shown in Fig. 3.7. Since Si is a group IV element, it needs four electrons to fill its shell. A P atom can pose as a Si atom and form four covalent bonds if it can somehow rid itself of one electron. So in fact with very little provocation, it will do so and “donate” an electron to the crystal. The group V element is thus an ionized donor atom. It is important to note that each dopant atom creates an electron, but not a hole. The ionized atom is an *immobile* positive charge. The material is said to become n-type.

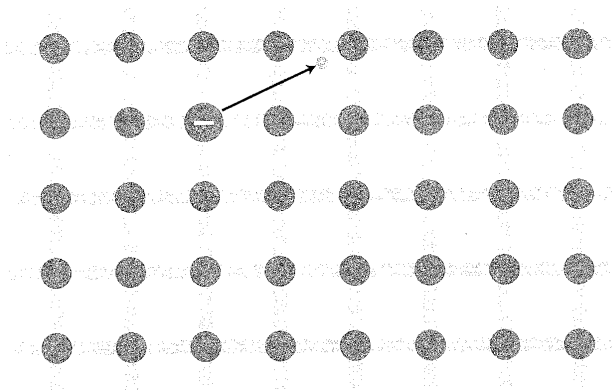


Figure 3.8 An acceptor dopant atom in the Si lattice contributes a free hole.

In contrast, a group III element, such as boron, is an “acceptor” dopant, as shown in Fig. 3.8. Since it has only three electrons in a four electron covalent bond world, it has a hole, so to speak. This hole is an energetically favorable state so other nearby valence electrons are likely to occupy this state, thus creating an ionized negative immobile charge site. In stealing an electron from the crystal it creates a hole that can travel through the crystal as discussed before. The material is said to become p-type.

In practice we actually dope a material with both acceptors and donors as we grow several layers of n-type and p-type material. The difference between the concentration of donors and acceptors will determine the n- or p-typeness of the material. It is also important to note that our arguments so far assume that the temperature T is large enough so that thermal energy is sufficient to ionize the dopants. If the temperature becomes too low, a process of freeze-out occurs, dramatically altering the conductivity of the material.

The motion of both electrons and holes in a crystal is governed by the band structure of the material. Fortunately, we can ignore these complexities and simply assign an effective mass to holes and electrons and treat them as free particles.

Acceptor and donor accounting

Let us begin by counting electrons in the donor case. The total charge density in the material is given by

$$\rho = \underbrace{-qn_0}_{\text{free electrons}} + \underbrace{qp_0}_{\text{free holes}} + \underbrace{qN_d}_{\text{immobile ions}} = 0 \quad (3.32)$$

By the law of mass action, in equilibrium, the product of electrons and holes is a constant, or $np = n_i^2(T)$. Substituting for the holes we have

$$-qn_0 + q\frac{n_i^2}{n_0} + qN_d = 0 \quad (3.33)$$

or

$$-n_0^2 + N_d n_0 + n_i^2 = 0 \quad (3.34)$$

Solving the quadratic equation, we arrive at the physical solution

$$n_0 = \frac{N_d + \sqrt{N_d^2 + 4n_i^2}}{2} \approx N_d \quad (3.35)$$

where we have assumed that $N_d \gg n_i$, since the doping $N_d \sim 10^{11} - 10^{18}$, usually several orders of magnitude larger than n_i . The same conclusion holds for acceptors, so that for a p-type material we have

$$p_0 \approx N_a \quad (3.36)$$

In the case of a compensated material,² it is the net doping that produces the free carriers

$$p_0 \approx N_a - N_d \quad (3.37)$$

Free carrier mobility

Since the conductivity of a material depends on the number of free carriers (Eq. (3.16))

$$\sigma = q\mu N \quad (3.38)$$

we have in effect a technique to produce a material of desired resistivity. If the material is n-type, this is simply

$$\sigma = q\mu_n N_d \quad (3.39)$$

where μ_n is the electron mobility. Unfortunately, the mobility is not a constant as we change the material doping. As we have remarked earlier, any disturbance to the periodic structure of a crystal acts as a scattering site and thus, not surprisingly, as we increase the doping, we decrease the mobility of free carriers. This is confirmed by the experimental plot of the bulk mobility of electrons and holes in Si, as shown in Fig. 3.9. As expected, the total doping concentration, $N_d + N_a$, affects the mobility, not the net doping, $|N_d - N_a|$.

Free carrier mobility model

The mobility of free carriers plays an important role in the design of devices and ultimately in the performance of circuits. It is therefore very important to produce an accurate model for the mobility. The equation

$$J = \sigma E = qn\mu_n E \quad (3.40)$$

is too good to be true. It implies that we can create current without bound by simply making E sufficiently large. In modern nanoscale semiconductor devices, we can in fact

² A compensated material is doped with both donors and acceptors impurities so that the overall characteristics are a function of the net doping.

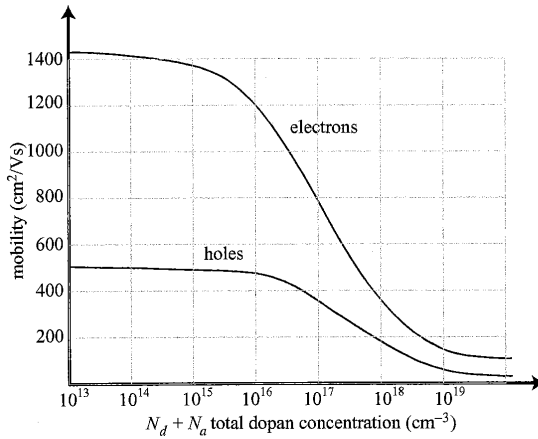


Figure 3.9 The mobility of electrons and holes as a function of total doping concentration.

create enormously large electric fields by simply applying a few volts on to our structures. Our various assumptions about the mobility of a free carrier begin to breakdown, though, under large field conditions. Most importantly, when creating the drift model of carrier current, we assumed that the drift velocity is a small fraction of the thermal velocity. As the fields are made very large, though, a linear mobility model predicts that we can in fact impart enough momentum to produce ultra high-speed carriers. Not only is the speed of light a limitation, but in practice we find that there is a limit due to increased phonon scattering at high fields, as shown in Fig. 3.10. We say that carriers become velocity saturated at high fields. This effect can be modeled by introducing a field dependent mobility model

$$\mu = \frac{\mu_0}{\left(1 + \left(\frac{E}{E_{sat}}\right)^n\right)^{1/n}} \quad (3.41)$$

where the coefficient n is fit to experimental data, with $n = 2$ for electrons and $n = 1$ for holes. The saturation velocity for electrons is about 8×10^6 cm/s and 6×10^6 cm/s for holes. From this we can solve for E_{sat}

$$v_{sat} = \mu E_{sat} = \mu_0 \frac{E_{sat}}{2^{1/n}} \quad (3.42)$$

There are other field dependencies in the mobility model. In particular, if current flow occurs near the surface of a semiconductor–oxide interface, such as in a MOSFET, then the mobility is found to be much lower at the surface. This is due to enhanced scattering sites due to the non-perfect interface between Si and SiO₂. A vertical field that tends to pull carriers towards the surface, therefore, will tend to lower the effective mobility.

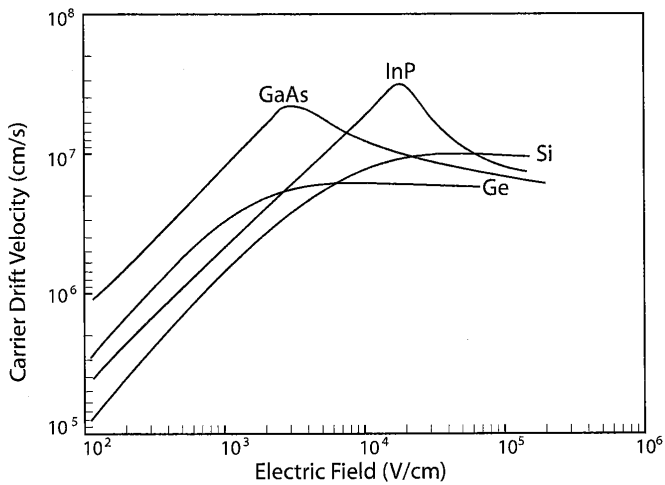


Figure 3.10 Velocity saturation for various materials. From [57].

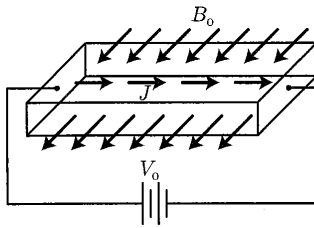


Figure 3.11 A rectangular conductor is placed inside a magnetic field.

On the other hand, in a MOSFET structure, we also find that the mobility drops at low vertical fields due to enhanced Coulomb scattering.³ This occurs since the field effect produces inversion, which masks the background dopant sites at the surface. At low fields, though, the dopants are exposed, causing enhanced scattering.

Hall effect

In this section we will review an important technique for measurement of free carrier density based on the Hall effect. Consider a conductor placed inside a magnetic field, as shown in Fig. 3.11. When current is traveling through a conductor, at any instant it experiences a

³ See Chapter 2 for a review of inversion in a MOSFET structure.

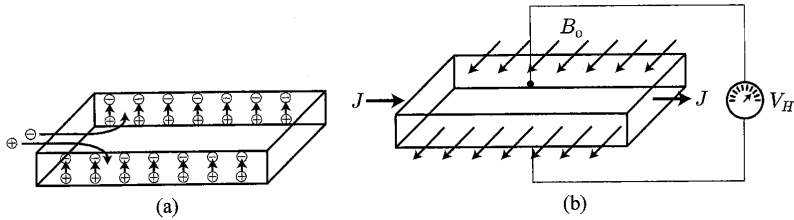


Figure 3.12 (a) Due to the magnetic force on free carriers, there is a charge pile-up on the top and bottom of the conductor. (b) This charge creates an internal electric field that opposes the movement of charges vertically.

force given by the Lorentz equation

$$\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \quad (3.43)$$

The force $q\mathbf{E}$ leads to conduction along the length of the bar (due to momentum relaxation) with average speed v_d but the magnetic field causes a downward deflection

$$\mathbf{F} = q\hat{x}E_0 - q\hat{y}v_d B_0 \quad (3.44)$$

In the steady state, the movement of charge down (or electrons up) creates an internal electric field which must balance the downward pull, as shown in Fig. 3.12. We expect a “Hall” voltage to develop across the top and bottom faces of the conducting bar

$$V_H = E_y d = v_d B_0 d \quad (3.45)$$

Since $v_d = \mu E_x$, and $J_x = \sigma E_x$, we can write $v_d = \mu J_x / \sigma$

$$V_H = \frac{\mu J_x}{\sigma} B_0 d \quad (3.46)$$

Since the conductivity of a material is given by $\sigma = qN\mu$, where q is the unit charge, N is the density of mobile charge carriers, and μ is the mobility of the carriers. Then

$$V_H = \frac{J_x B_0 d}{qN} \quad (3.47)$$

or

$$N = \frac{J_x B_0 d}{q V_H} = \frac{I B_0 d}{A q V_H} \quad (3.48)$$

Notice that all the quantities on the right-hand side are either known or easily measured. Thus the density of carriers can be measured indirectly through measuring the Hall voltage.

3.3 Diffusion

The process of diffusion is quite intuitive and well known. In Fig. 3.13, imagine that we fill the left chamber with a gas at temperature T . If we suddenly remove the divider, what happens? The gas will fill the entire volume of the new chamber.

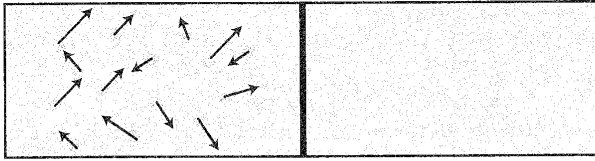


Figure 3.13 A chamber is separated into two sections by a partition. In the left section there is a gas at temperature T . If the partition is removed, we know by experience that the gas on the left will fill the entire chamber.

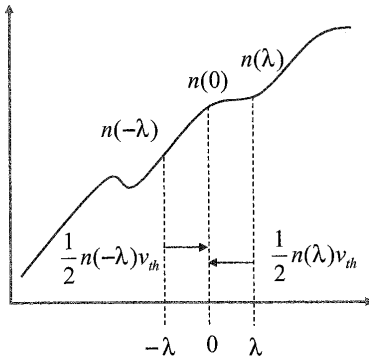


Figure 3.14 A concentration variation of free particles results in a flux proportional to the gradient of the concentration.

Whenever there exists a concentration gradient, we know from experience that over time the region of high concentration and low concentration balance as particles from the high concentration region move to the region of low concentration. How does this actually happen so naturally?

If each particle moves on average left or right, then eventually half will be in the right chamber. If the molecules were charged (or electrons), then there would be a net current flow. The diffusion current flows from high concentration to low concentration.

Let us quantify the diffusion current. Assume that the mean free path is λ . Find the flux of carriers crossing the $x = 0$ plane in Fig. 3.14

$$F = \frac{1}{2} v_{th} (n(-\lambda) - n(\lambda)) \quad (3.49)$$

where we assume that particles are moving with velocity v_{th} . The factor of $1/2$ accounts for the fact that half of the particles will be moving to the right into the $x = 0$ plane. If there are more particles at $x = -\lambda$ than $x = +\lambda$, then a net flux of particles will cross the $x = 0$ plane to the right. Expanding $n(x)$ about $x = 0$, we have

$$F = \frac{1}{2} v_{th} \left(\left[n(0) - \lambda \frac{dn}{dx} \right] - \left[n(0) + \lambda \frac{dn}{dx} \right] \right) \quad (3.50)$$

where we assume that the mean free path λ is small relative to the variation in the concentration. Simplifying we have

$$F = -v_{th}\lambda \frac{dn}{dx} \quad (3.51)$$

As expected, if $dn/dx > 0$, the net flux should be to the left. If the particles are electrons, then the current produced by this flux is given by

$$J = -qF = qv_{th}\lambda \frac{dn}{dx} \quad (3.52)$$

Since the thermal energy per degree of freedom is given by $kT/2$, we have

$$\frac{1}{2}m_n^*v_{th}^2 = \frac{1}{2}kT \quad (3.53)$$

The mean free path is related to the mean free time, or the average time between collisions, by $\lambda = v_{th}\tau_c$. We are also careful to use the effective mass m_n^* for electrons in a crystal. Manipulating our equations a bit

$$v_{th}\lambda = v_{th}^2\tau_c = kT \frac{\tau_c}{m_n^*} = \frac{kT}{q} \left(\frac{q\tau_c}{m_n^*} \right) \quad (3.54)$$

The parenthetical term is just the mobility for free carriers

$$J = qv_{th}\lambda \frac{dn}{dx} = q \left(\frac{kT}{q} \mu_n \right) \frac{dn}{dx} = qD_n \frac{dn}{dx} \quad (3.55)$$

The constant D_n is known as the diffusion constant and we see that it is related to the mobility. This relationship is known as the Einstein relation

$$D_n = \left(\frac{kT}{q} \right) \mu_n \quad (3.56)$$

which you will never forget, because it rhymes like a simple poem, *kay-tee over q equals dee over mu*.

3.4 Thermal noise

With all the random motion in a typical conductor, you would not be surprised to hear that every resistor is a source of noise. This noise, often referred to as Johnson noise, Nyquist noise, or simply thermal noise, is present in every conductor and linearly proportional to the temperature T . The key point about the noise is that it has zero average value but non-zero RMS, with a noise voltage as shown in Fig. 3.15a. At any given instant, if we sample the noise voltage, it has a Gaussian amplitude distribution. The mathematical description of such a function is not an easy task. We will avoid many of the complicated issues and just work with RMS values.

Johnson was the first to observe this noise experimentally. Nyquist was probably the first to explain it. We will also demonstrate that this form of noise is consistent with the theory of thermodynamics.

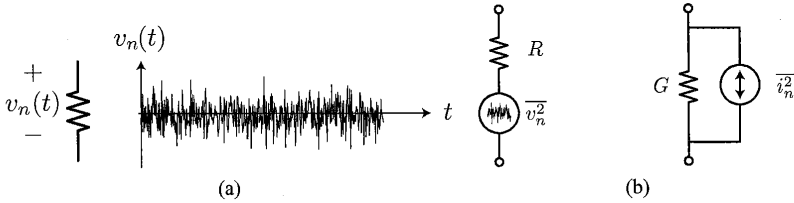


Figure 3.15 (a) The noise voltage of a resistor as a function of time. (b) Equivalent circuit representation as a noiseless resistor in series or shunt with a noise voltage or current.

The average value of the noise waveform is zero

$$\overline{v_n(t)} = \langle v_n(t) \rangle = \frac{1}{T} \int_T v_n(t) dt = 0 \quad (3.57)$$

The mean is also zero if we freeze time and take an infinite number of samples from identical resistors. The variance, though, is non-zero. Equivalently, we may say that the signal power is non-zero

$$\overline{v_n(t)^2} = \frac{1}{T} \int_T v_n^2(t) dt \neq 0 \quad (3.58)$$

The RMS (root-mean-square) voltage is given by

$$v_{n,rms} = \sqrt{\overline{v_n(t)^2}} \quad (3.59)$$

The power spectrum of the noise shows the concentration of noise power at any given frequency. Many noise sources are “white” in that the spectrum is flat (up to extremely high frequencies). In such cases the noise waveform is totally unpredictable as a function of time. In other words, there is absolutely no correlation between the noise waveform at time t_1 and some later time $t_1 + \delta t$, no matter how small we make δt .

As shown in Fig. 3.15b, we can represent the noise as a voltage source in series with the resistor of the RMS value given by

$$\overline{v_R^2} = 4kTRB \quad (3.60)$$

where B is the bandwidth of observation. Equivalently, we can represent this with a shunt current source

$$\overline{i_n^2} = 4kTGB \quad (3.61)$$

Here B is the bandwidth of observation and kT is Boltzmann’s constant times the temperature of observation. This result comes from thermodynamic considerations, thus explaining the appearance of kT . Often we speak of the “spot noise,” or the noise in a specific narrow-band δf

$$\overline{v_n^2} = 4kTR\delta f \quad (3.62)$$

Since the noise is white, the shape of the noise spectrum is determined by the external elements (L s and C s).

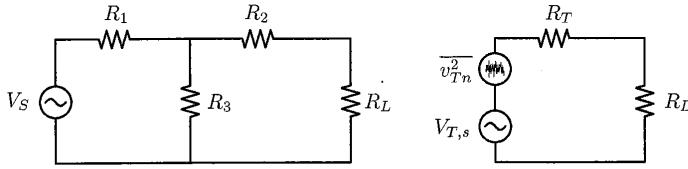


Figure 3.16 Any arbitrary resistive circuit can be reduced to a single equivalent noise source.

Example 4

Suppose that $R = 10\text{k}\Omega$ and $T = 20^\circ\text{C} = 293\text{K}$.

$$4kT = 1.62 \times 10^{-20} \quad (3.63)$$

$$\overline{v_n^2} = 1.62 \times 10^{-16} \times B \quad (3.64)$$

$$v_{n,rms} = \sqrt{\overline{v_n(t)^2}} = 1.27 \times 10^{-8} \sqrt{B} \quad (3.65)$$

If we limit the bandwidth of observation to $B = 10^6\text{MHz}$, then we have

$$v_{n,rms} \approx 13\mu\text{V} \quad (3.66)$$

This represents the limit for the smallest voltage we can resolve across this resistor in this bandwidth.

If we put two resistors in series, then the mean square noise voltage is given by

$$\overline{v_n^2} = 4kT(R_1 + R_2)B = \overline{v_{n1}^2} + \overline{v_{n2}^2} \quad (3.67)$$

We add noise powers, *not* the noise voltages. Likewise, for two resistors in parallel, we can add the mean square currents

$$\overline{i_n^2} = 4kT(G_1 + G_2)B = \overline{i_{n1}^2} + \overline{i_{n2}^2} \quad (3.68)$$

This holds for any pair of independent noise sources (zero correlation).

For an arbitrary resistive circuit, such as Fig. 3.16, we can find the equivalent noise by using a Thevenin (Norton) equivalent circuit or by transforming all noise sources to the output by the appropriate *power* gain (e.g. voltage squared or current squared)

$$V_{T,s} = V_S \frac{R_3}{R_1 + R_3} \quad (3.69)$$

$$\overline{v_{Tn}^2} = 4kT R_T B = 4kT (R_2 + R_1 || R_3) B \quad (3.70)$$

For a general linear circuit, such as the black box shown in Fig. 3.17, the mean square noise voltage (current) at any port is given by the equivalent input resistance (conductance)

$$\overline{v_{eq}^2} = 4kT \Re(Z(f)) \delta f \quad (3.71)$$

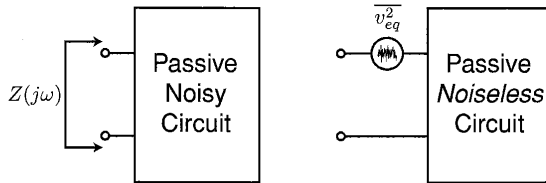


Figure 3.17 The equivalent noise of an arbitrary linear one-port circuit.

This is the “spot” noise. If the network has a filtering property, then we integrate over the band of interest

$$\overline{v_{T,eq}^2} = 4kT \int_B \Re(Z(f))df \quad (3.72)$$

Unlike resistors, ideal (lossless) L s and C s *do not* generate noise. They do shape the noise due to their frequency dependence.

Example 5

Noise of an RC circuit

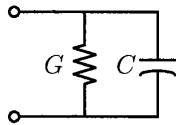


Figure 3.18 An example RC circuit for calculation of the total thermal noise.

To find the equivalent mean square noise voltage of an RC circuit, begin by calculating the impedance

$$Z = \frac{1}{Y} = \frac{1}{G + j\omega C} = \frac{G - j\omega C}{G^2 + \omega^2 C^2} \quad (3.73)$$

Integrating the noise over all frequencies, we have

$$\overline{v_n^2} = \frac{4kT}{2\pi} \int_0^\infty \frac{G}{G^2 + \omega^2 C^2} d\omega = \frac{kT}{C} \quad (3.74)$$

Notice the result is *independent* of R . Since the noise/bandwidth is proportional/inversely proportional to R , its influence cancels out.

Thermodynamic origin of noise

As shown in Fig. 3.19, imagine a black body held at temperature T . Inside this body we insert an antenna with characteristic impedance $Z_0 = R$ and through a small opening, we

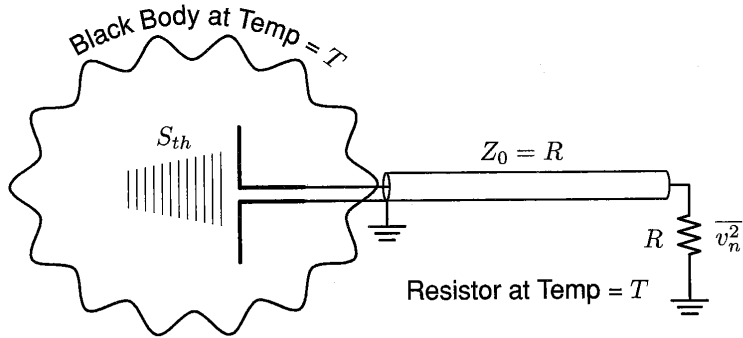


Figure 3.19 A black body held at temperature T exchanges thermal energy with a termination resistor R .

route a transmission line to a termination resistor R , also held at a temperature T , with a small opening where we insert a transmission line.

From thermodynamic considerations, we have the black body radiation density

$$S = \frac{2kT}{\lambda^2} \quad (3.75)$$

Our antenna will capture a fraction of this energy $A_e S$, where $A_e = \frac{\lambda^2}{4\pi}$ is the antenna effective cross. If we integrate over the solid angle, we have

$$E_L = \frac{1}{2} S A_e 4\pi \quad (3.76)$$

where only half of the energy is captured by the antenna due to the random polarization of the incoming radiation. Substituting in the above equation, we have the important results

$$E_L = \frac{1}{2} \frac{2kT}{\lambda^2} \frac{\lambda^2}{4\pi} 4\pi = kT \quad (3.77)$$

If we observe a bandwidth B , we capture an amount of power kTB . So we see that the resistor captures kTB Watts of power from the black body. But since the black body is in thermodynamic equilibrium with the resistor, the resistor must likewise deliver the same power back to the black body!

Now let us imagine that the resistor is a noiseless resistor in series with a voltage source. Under matched conditions, the RMS power delivered to a load is given by

$$P_L = \frac{\left(\frac{v_n}{2}\right)^2}{R} = \frac{\overline{v^2}}{4R} \quad (3.78)$$

If we equate this with the power delivered by a noise resistor

$$kT\delta f = \frac{\overline{v^2}}{4R} \quad (3.79)$$

or

$$\overline{v^2} = 4kTR\delta f \quad (3.80)$$

which is in fact the physically observed amount of noise generated by a resistor.

3.5 References

A lot of this material was put together while teaching a device physics undergraduate course at UC Berkeley and thus some of the material is drawn from lecture notes put together by Howe and Sodini [27]. The discussion on the conductivity of a gas and metal borrows heavily from Purcell [49]. Purcell's book on electromagnetics is wonderfully good and very accessible. The presentation from thermodynamic origin of noise comes from Hagen's excellent book [26]. This book is a very good source for learning the fundamentals of RF circuits, presented for a general audience but with a lot of great insight for even the expert.

4 Ampère, Faraday, and Maxwell

4.1 Ampère: static magnetic fields

Experimental observations

Consider a pair of parallel wires carrying steady currents I_1 and I_2 . Since steady currents imply zero net charge distribution, there should be no electrostatic force between these current-carrying wires. But experimentally we do observe a force which tends to be attractive if the currents are in the same direction and repulsive if the currents are in the opposite direction. This new force is in fact an electrostatic force if we consider the problem from a relativistic point of view! Even though the net charge on each current-carrying conductor is zero in a static reference frame, in a moving reference frame there is net charge density and hence a force exists.

Magnetic force

Through careful observations, Ampère demonstrated that this force can be computed using the following equation

$$dF_m = \frac{\mu_0}{4\pi} \frac{I_2 d\ell_2 \times I_1 d\ell_1 \times \hat{\mathbf{R}}}{R^2} \quad (4.1)$$

The resemblance to the Coulomb force equation is notable. Both forces fall like $1/R^2$. For steady currents, $\nabla \cdot \mathbf{J} = 0$ implies that the currents must flow in loops. Thus we can calculate the force between two loops as follows

$$F_m = \oint_{C_1} \oint_{C_2} \frac{\mu_0}{4\pi} \frac{I_2 d\ell_2 \times I_1 d\ell_1 \times \hat{\mathbf{R}}}{R^2} \quad (4.2)$$

Magnetic field

Just as in the case of electric forces, the concept of “action at a distance” is disturbing and counterintuitive. Thus we prefer to think of the current in loop C_1 generating a “field” and then we say that this field interacts with the current in loop C_2 to generate a force. Simply reordering the magnetic force equation gives

$$F_m = \oint_{C_2} I_2 d\ell_2 \times \underbrace{\frac{\mu_0}{4\pi} \oint_{C_1} \frac{I_1 d\ell_1 \times \hat{\mathbf{R}}}{R^2}}_{\mathbf{B}} \quad (4.3)$$

where we have interpreted the second term as a magnetic field $\mathbf{B}$. Loop C_2 is the source of the field and loop C_1 is the field point under consideration. The unit of $\mathbf{B}$ is the Tesla (T), where $1\text{T} = 10^4\text{G}$, in terms of the CGS units of Gauss (G).

Units of magnetic field

The Tesla (T) and Gauss (G) are derived units. Since $F \propto I^2 \mu$, the units of μ are simply $\text{N} \cdot \text{A}^{-2}$. This is more commonly known as $\text{H} \cdot \text{m}^{-1}$. The units of the magnetic field are therefore

$$[B] = [\mu] \text{A} \cdot \text{m} \cdot \text{m}^{-2} = \text{H} \cdot \text{A} \cdot \text{m}^{-2} \quad (4.4)$$

Note that the units of D are $\text{C} \cdot \text{m}^{-2}$, which can be written as $\text{F} \cdot \text{V} \cdot \text{m}^{-2}$. From circuit theory we know that voltage is proportional to ωLI , so LI has units of $\frac{\text{V}}{\omega}$. So the unit of $[B]$ is $\text{V} \cdot \text{s} \cdot \text{m}^{-2}$. For reference, the magnetic field of the earth is only .5G, so 1T is a very large field.

Direction of magnetic force

Due to the vector cross product, the direction of the force of the magnetic field is perpendicular to the direction of motion and the magnetic field. Use the right-hand rule to figure out the direction of $\mathbf{F}_m$ in any given situation.

E and B duality

For a point charge dq , the electric force is given by

$$\mathbf{F}_e = q\mathbf{E} \quad (4.5)$$

For the magnetic force for a point charge in a current loop, we have

$$\mathbf{F}_m = Id\boldsymbol{\ell} \times \mathbf{B} = qN d\boldsymbol{\ell} \mathbf{v} \times \mathbf{B} \quad (4.6)$$

The equations for $\mathbf{E}$ and $\mathbf{B}$ are also similar when we consider an arbitrary current density $\mathbf{J}$ and charge density ρ

$$\mathbf{E} = \frac{1}{4\pi\epsilon} \int_V \frac{\rho(\mathbf{r}')\mathbf{R}}{R^2} dV' \quad (4.7)$$

$$\mathbf{B} = \frac{1}{4\pi\mu^{-1}} \int_V \frac{\mathbf{J}(\mathbf{r}') \times \hat{\mathbf{R}}}{R^2} dV' \quad (4.8)$$

Magnetic charge

We may now compare the magnetic field to the electric field and look for similarity and differences. In this book we shall not discuss the relativistic viewpoint that explains the link between electric and the magnetic field. Instead, we shall assume that the magnetic field is an entity of its own. Apparently, the source of magnetic field is moving charge (currents),

whereas the source of electric fields is charge. But what about magnetic charges? Is there any reason to believe that nature should be asymmetric and give us electrical charge and not magnetic charge? If magnetic charge existed, then the argument for Gauss' Law would apply

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = Q_m \quad (4.9)$$

where Q_m is the amount of magnetic charge inside the volume V bounded by surface S . But no one has ever observed any magnetic charge! So, for all practical purposes, we can assume that $Q_m = 0$ and so Gauss' Law applied to magnetic fields yields

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0 \quad (4.10)$$

Divergence of B

By the Divergence Theorem, locally this relation translates into

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{B} dV = 0 \quad (4.11)$$

Since this is true for any surface S , the integrand must be identically zero

$$\nabla \cdot \mathbf{B} = 0 \quad (4.12)$$

A vector field with zero divergence is known as a solenoidal field. We have already encountered such a field since $\nabla \cdot \mathbf{J} = 0$. Such a field does not have any sources and thus always curls back onto itself. $\mathbf{B}$ fields are thus always loops.

Divergence of curl

Let's calculate the divergence of the curl of an arbitrary vector field $\mathbf{A}$, $\nabla \cdot \nabla \times \mathbf{A}$. Let's compute the volume of the above quantity and apply the Divergence Theorem

$$\int_V \nabla \cdot \nabla \times \mathbf{A} dV = \oint_S \nabla \times \mathbf{A} \cdot d\mathbf{S} \quad (4.13)$$

To compute the surface integral, consider a new surface S' with a hole in it. The surface integral of $\nabla \times \mathbf{A}$ can be written as the line integral using Stoke's Theorem

$$\int_{S'} \nabla \times \mathbf{A} \cdot d\mathbf{S} = \oint_C \mathbf{A} \cdot d\mathbf{l} \quad (4.14)$$

where C is the perimeter of the hole. As we shrink this hole to a point, the right-hand side goes to zero and the surface integral turns into the closed-surface integral. Thus

$$\int_V \nabla \cdot \nabla \times \mathbf{A} dV = \oint_S \nabla \times \mathbf{A} \cdot d\mathbf{S} = 0 \quad (4.15)$$

Since this is true for any volume V , it must be that

$$\nabla \cdot \nabla \times \mathbf{A} = 0 \quad (4.16)$$

Thus a solenoidal vector can always be written as the curl of another vector. Thus the magnetic field $\mathbf{B}$ can be written as

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (4.17)$$

Ampère's Law

One of the fundamental relations for the magnetic field is Ampère's Law. It is analogous to Gauss' Law. We can derive it by taking the curl of the magnetic field

$$\nabla \times \mathbf{B} = \nabla \times \frac{1}{4\pi\mu^{-1}} \int_V \frac{\mathbf{J}(\mathbf{r}') \times \hat{\mathbf{R}}}{R^2} dV' \quad (4.18)$$

After some *painful* manipulations, this can be simplified to the famous Ampère's Law

$$\nabla \times \mathbf{B} = \mu \mathbf{J} \quad (4.19)$$

Now apply Stoke's Theorem

$$\int_S \nabla \times \mathbf{B} \cdot d\mathbf{S} = \oint_C \mathbf{B} \cdot d\mathbf{l} = \int_S \mu \mathbf{J} \cdot d\mathbf{S} = \mu I \quad (4.20)$$

Application of Ampère's Law

Ampère's Law is very handy in situations involving cylindrical symmetry. This is analogous to applying Gauss' Law to problems with spherical symmetry. For example, consider the magnetic field due to a long wire. The field should have no r or z dependence (by symmetry) so the integral of B over a circle enclosing the wire is simply a constant times the perimeter

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = 2\pi r B_r = \mu I \quad (4.21)$$

So the magnetic field drops as $1/r$

$$B(r) = \frac{\mu I}{2\pi r} \quad (4.22)$$

Example 6

Consider the magnetic field of a two-dimensional round wire of radius a carrying a uniform current I shown in Fig. 4.1a. From symmetry we deduce that the magnetic field is circumferential. Invoking Ampère's Law, the closed line integral of the field around the wire is equal to the current enclosed by the loop. Let's take a loop inside the wire so that $r < a$

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \int_0^{2\pi} B_\theta d\theta = 2\pi r B_\theta(r) = \mu I \frac{\pi r^2}{\pi a^2} \quad (4.23)$$

since $B(r) = B_\theta(r)$ is a constant as a function of θ , and a circle of radius $r < a$ encloses a fraction r^2/a^2 of the surface area of the wire. Hence the magnetic field

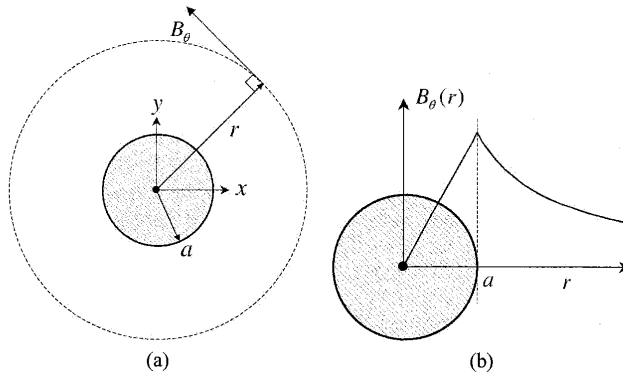


Figure 4.1 (a) A two-dimensional round wire carrying a current I out of the surface of the page. (b) The radial dependence of the magnetic field inside and outside of the wire.

inside the wire increases linearly as a function of r

$$B_{\theta}(r) = \frac{\mu I r}{2\pi a^2} \quad (4.24)$$

Outside the wire, $r > a$, we enclose the entire current and so

$$B_{\theta}(r) = \frac{\mu I}{2\pi r} \quad (4.25)$$

The functional dependence of $B_{\theta}(r)$ is plotted in Fig. 4.1b.

Example 7

Consider the magnetic field between two infinitely long wires carrying oppositely directed currents. The geometry is shown in Fig. 4.2a where a positive current coming out of the plane of the page is shown at $x = -d$ and a negative current into the plane of the page is at $x = +d$. We will solve this problem by invoking symmetry and superposition. From symmetry it is clear that the magnetic field of a single filament is circularly directed. From Ampère's Law, the closed line integral of the field around the wire is equal to the current enclosed by the loop

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \int_0^{2\pi} B_{\theta} d\theta = 2\pi r B_{\theta}(r) = \mu I \quad (4.26)$$

since $B(r) = B_{\theta}(r)$ is a constant as a function of θ . Thus we have the result that

$$B_{\theta} = \frac{\mu I}{2\pi r} \quad (4.27)$$

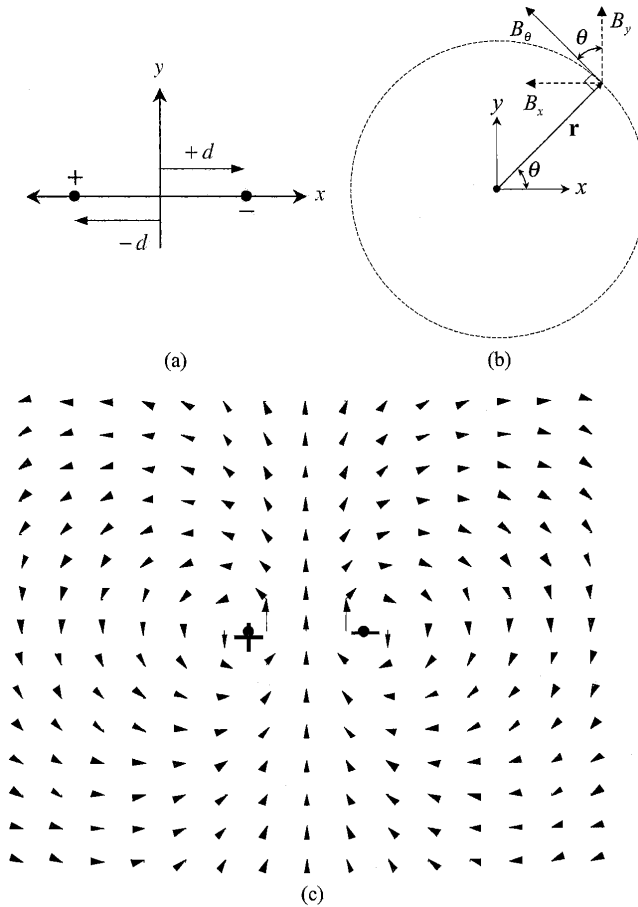


Figure 4.2 (a) Two filaments carrying oppositely directed currents. (b) The magnetic field of a single filament. (c) The magnetic field of the two filament system.

Writing the above result in Cartesian coordinates (see Fig. 4.2b for the geometry) we have

$$\begin{aligned} B_x &= -B_\theta \sin \theta = \frac{-\mu I y}{2\pi(x^2 + y^2)} \\ B_y &= B_\theta \cos \theta = \frac{\mu I x}{2\pi(x^2 + y^2)} \end{aligned} \quad (4.28)$$

By superposition the field for two filaments is simply the vectorial summation of the field of the individual filaments displaced to $x = d$ and $x = -d$

$$\begin{aligned} B_x &= \frac{-\mu I y}{2\pi ((x+d)^2 + y^2)} + \frac{\mu I y}{2\pi ((x-d)^2 + y^2)} \\ B_y &= \frac{-\mu I (x+d)}{2\pi ((x+d)^2 + y^2)} - \frac{\mu I (x-d)}{2\pi ((x-d)^2 + y^2)} \end{aligned} \quad (4.29)$$

This field is plotted in Fig. 4.2c using the following commands in *Mathematica*

```
In[1]:= Bx = -y/((x+1)^2+y^2) + y/((x-1)^2+y^2)
In[2]:= By = (x+1)/((x+1)^2+y^2) - (x-1)/((x-1)^2+y^2)
In[3]:= B = {Bx,By};
In[4]:= <<Graphics'PlotField'
In[5]:= bfield = PlotVectorField[B,{x,-5,5},{y,-4,4}];
In[6]:= cur1 = Point[{-1,0}];
In[7]:= cur2 = Point[{1,0}];
In[8]:= cur = {cur1,cur2};
In[9]:= Show[bfield,Graphics[{PointSize[.02],cur}]] }
```

Magnetic vector potential

Earlier we showed that we can also define a vector $\mathbf{A}$ such that $\mathbf{B} = \nabla \times \mathbf{A}$. Since $\nabla \times \mathbf{B} = \mu \mathbf{J}$

$$\nabla \times \nabla \times \mathbf{A} = \mu \mathbf{J} \quad (4.30)$$

We can apply the vector identity for $\nabla \times \nabla \times \mathbf{A}$

$$\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} \quad (4.31)$$

Given that the divergence of $\mathbf{A}$ is arbitrary, let us choose the most convenient value. In magnetostatics that is $\nabla \cdot \mathbf{A} = 0$. Then we have

$$\nabla^2 \mathbf{A} = -\mu \mathbf{J} \quad (4.32)$$

We have met this equation before!

Equations for potential

The vector Laplacian can be written as three scalar Laplacian equations (using rectangular coordinates). For instance, the x-component is given by

$$\nabla^2 A_x = -\mu J_x \quad (4.33)$$

By analogy with the scalar potential, therefore, the solution is given by

$$A_x = \frac{\mu}{4\pi} \int_V \frac{J_x(\mathbf{r}') dV'}{|\mathbf{r} - \mathbf{r}'|} \quad (4.34)$$

and, in general, the total vector potential is given by

$$\mathbf{A} = \frac{\mu}{4\pi} \int_V \frac{\mathbf{J}(\mathbf{r}') dV'}{|\mathbf{r} - \mathbf{r}'|} \quad (4.35)$$

Why use vector potential?

The vector potential $\mathbf{A}$ is easier to calculate than $\mathbf{B}$ since each component is a simple scalar calculation. Also, the direction of $\mathbf{A}$ is easy to determine since it follows $\mathbf{J}$. At this point the vector potential seems like a mathematical creation and it does not seem to have any physical relevance. This is compounded by the fact that $\nabla \cdot \mathbf{A}$ is arbitrary. Later on, we will see that $\mathbf{A}$ has a lot of physical relevance and in some ways it is more fundamental than the vector $\mathbf{B}$.

From vector $\mathbf{A}$ to $\mathbf{B}$

Now that we have an equation for $\mathbf{A}$, we can verify that it is indeed consistent with the experimentally observed equation for $\mathbf{B}$

$$\nabla \times \mathbf{A} = \nabla \times \frac{\mu}{4\pi} \int_V \frac{\mathbf{J}(\mathbf{r}') dV'}{|\mathbf{r} - \mathbf{r}'|} \quad (4.36)$$

Interchanging the order of integration and differentiation

$$\nabla \times \mathbf{A} = \frac{\mu}{4\pi} \int_V \nabla \times \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dV' \quad (4.37)$$

We now need some fancy footwork to go further.

Yet another vector identify

It is relatively easy to show that for a scalar function ϕ times a vector field $\mathbf{F}$

$$\nabla \times \phi \mathbf{F} = \nabla \phi \times \mathbf{F} + \phi \nabla \times \mathbf{F} \quad (4.38)$$

Applying this to our case, note that ∇ operates on the coordinates $\mathbf{r}$ whereas $\mathbf{J}(\mathbf{r}')$ is a function of the primed coordinates, and hence a constant

$$\nabla \times \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} = \nabla \frac{1}{|\mathbf{r} - \mathbf{r}'|} \times \mathbf{J}(\mathbf{r}') = -\frac{\hat{\mathbf{R}}}{|\mathbf{r} - \mathbf{r}'|^2} \times \mathbf{J}(\mathbf{r}') \quad (4.39)$$

where we have used

$$\nabla \frac{1}{|\mathbf{r} - \mathbf{r}'|} = \frac{\hat{\mathbf{R}}}{|\mathbf{r} - \mathbf{r}'|^2} \quad (4.40)$$

The expression for B matches up with the experimentally observed equation

$$\mathbf{B} = \nabla \times \mathbf{A} = \frac{\mu}{4\pi} \int_V \frac{\mathbf{J}(\mathbf{r}') \times \hat{\mathbf{R}}}{R^2} dV' \quad (4.41)$$

4.2 Magnetic materials

Magnetization vector

We would like to study magnetic fields in magnetic materials. Let's define the magnetization vector as

$$\mathbf{M} = \lim_{\Delta V \rightarrow 0} \frac{\sum_k \mathbf{m}_k}{\Delta V} \quad (4.42)$$

where $\mathbf{m}_k$ is the magnetic dipole of an atom or molecule. The vector potential due to these magnetic dipoles in a differential volume dV' is given by

$$d\mathbf{A} = \mu_0 \frac{\mathbf{M} \times \hat{\mathbf{r}}}{4\pi R^2} dV' \quad (4.43)$$

so

$$\mathbf{A} = \frac{\mu_0}{4\pi} \int_{V'} \frac{\mathbf{M} \times \hat{\mathbf{r}}}{R^2} dV' \quad (4.44)$$

Using

$$\nabla' \left(\frac{1}{R} \right) = \frac{\hat{\mathbf{r}}}{R^2} \quad (4.45)$$

$$\mathbf{A} = \frac{\mu_0}{4\pi} \int_{V'} \mathbf{M} \times \nabla' \left(\frac{1}{R} \right) dV' \quad (4.46)$$

Consider the vector identity

$$\nabla' \times \left(\frac{\mathbf{M}}{R} \right) = \frac{1}{R} \nabla' \times \mathbf{M} + \nabla' \left(\frac{1}{R} \right) \times \mathbf{M} \quad (4.47)$$

We can thus break the vector potential into two terms

$$\mathbf{A} = \frac{\mu_0}{4\pi} \int_{V'} \frac{\nabla' \times \mathbf{M}}{R} dV' - \frac{\mu_0}{4\pi} \int_{V'} \nabla' \times \left(\frac{\mathbf{M}}{R} \right) dV' \quad (4.48)$$

Another Divergence Theorem

Consider the vector $\mathbf{u} = \mathbf{a} \times \mathbf{v}$, where $\mathbf{a}$ is an arbitrary constant. Then

$$\nabla \cdot \mathbf{u} = \nabla \cdot (\mathbf{a} \times \mathbf{v}) = (\nabla \times \mathbf{a}) \cdot \mathbf{v} - (\nabla \times \mathbf{v}) \cdot \mathbf{a} = -(\nabla \times \mathbf{v}) \cdot \mathbf{a} \quad (4.49)$$

Now apply the Divergence Theorem to $\nabla \cdot \mathbf{u}$

$$\int_V -(\nabla \times \mathbf{v}) \cdot \mathbf{a} dV = \oint_S (\mathbf{a} \times \mathbf{v}) \cdot d\mathbf{S} = \oint_S ((\mathbf{a} \times \mathbf{v}) \cdot \hat{\mathbf{n}}) dS \quad (4.50)$$

where $\hat{\mathbf{n}}$ is a unit vector perpendicular to the surface. Reordering the vector triple product

$$\oint_S (\mathbf{a} \cdot \mathbf{v} \times \hat{\mathbf{n}}) dS \quad (4.51)$$

Since the vector $\mathbf{a}$ is constant, we can pull it out of the integrals

$$\mathbf{a} \cdot \int_V (-\nabla \times \mathbf{v}) dV = \mathbf{a} \cdot \oint_S \mathbf{v} \times \hat{\mathbf{n}} dS \quad (4.52)$$

The vector $\mathbf{a}$ is arbitrary, so we have

$$\int_V (\nabla \times \mathbf{v}) dV = - \oint_S \mathbf{v} \times \hat{\mathbf{n}} dS \quad (4.53)$$

Applying this to the second term of the vector potential

$$\int_{V'} \nabla' \times \left(\frac{\mathbf{M}}{R} \right) dv' = - \oint_S \frac{\mathbf{M} \times \hat{\mathbf{n}}}{R} dS \quad (4.54)$$

Vector potential due to magnetization

The vector potential due to magnetization has a volume component and a surface component

$$\mathbf{A} = \frac{\mu_0}{4\pi} \int_{V'} \frac{\nabla' \times \mathbf{M}}{R} dv' + \frac{\mu_0}{4\pi} \oint_S \frac{\mathbf{M} \times \hat{\mathbf{n}}}{R} \cdot d\mathbf{S} dv' \quad (4.55)$$

We can thus define an equivalent magnetic volume current density

$$\mathbf{J}_m = \nabla \times \mathbf{M} \quad (4.56)$$

and an equivalent magnetic surface current density

$$\mathbf{J}_s = \mathbf{M} \times \hat{\mathbf{n}} \quad (4.57)$$

Volume and surface currents

In Fig. 4.3, we can see that for uniform magnetization, all the internal currents cancel and only the magnetization vector on the boundary (surface) contributes to the integral.

Relative permeability

We can include the effects of materials on the macroscopic magnetic field by including a volume current $\nabla \times \mathbf{M}$ in Ampère's equation

$$\nabla \times \mathbf{B} = \mu_0(\mathbf{J} + \nabla \times \mathbf{M}) \quad (4.58)$$

or

$$\nabla \times \left(\frac{\mathbf{B}}{\mu_0} - \mathbf{M} \right) = \mathbf{J} \quad (4.59)$$

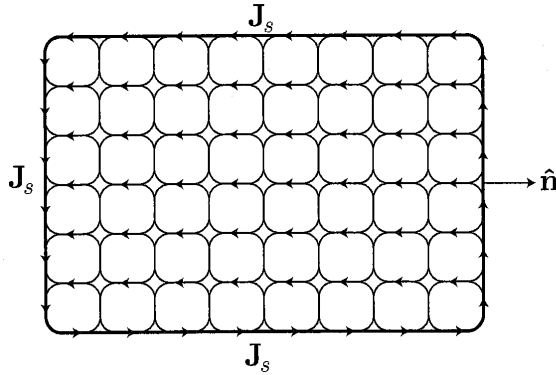


Figure 4.3 The magnetization current flowing in a uniform conductor.

We thus have defined a new quantity $\mathbf{H}$

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} \quad (4.60)$$

The units of $\mathbf{H}$, the *magnetic field*, are A/m.

Ampère's equation for media

We can thus state that for any medium under static conditions

$$\nabla \times \mathbf{H} = \mathbf{J} \quad (4.61)$$

equivalently

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = I \quad (4.62)$$

Since linear materials respond to the external field in a linear fashion

$$\mathbf{M} = \chi_m \mathbf{H} \quad (4.63)$$

so

$$\mathbf{B} = \mu(1 + \chi_m)\mathbf{H} = \mu\mathbf{H} \quad (4.64)$$

or

$$\mathbf{H} = \frac{1}{\mu} \mathbf{B} \quad (4.65)$$

The table in Fig. 4.4 can serve as a memory aid as it shows the duality between the electric and magnetic field.¹

¹ I personally don't like this choice since to me $\mathbf{E}$ and $\mathbf{B}$ are "real" and so the equations should be arranged to magnify this analogy. Unfortunately, the equations are not organized this way (partly due to choice of units) so we'll stick with convention.

E	H	ρ	J
D	B	V	A
ϵ	μ^{-1}	$\cdot$	$\times$
P	M	$\times$	$\cdot$

Figure 4.4 The duality between electric and magnetic fields.

Magnetic materials

Magnetic materials are classified as follows:

- Diamagnetic: $\mu_r \leq 1$, usually χ_m is a small negative number
- Paramagnetic: $\mu_r \geq 1$, usually χ_m is a small positive number
- Ferromagnetic: $\mu_r \gg 1$, thus χ_m is a large positive number

Most materials in nature are diamagnetic. To fully understand the magnetic behavior of materials requires a detailed study (and quantum mechanics). In this book we mostly assume $\mu \approx \mu_0$.

The magnetic flux density **B** is related to the magnetic field **H** by

$$\mathbf{B} = \mu \mathbf{H} = \mu_r \mu_0 \mathbf{H}$$

$$\mu_0 = 4\pi \times 10^{-7} \text{ H/m}$$

The relative permeability μ_r measures the effect of constituent atomic and/or molecular magnetic dipole moments. Most materials in nature are diamagnetic. The induced magnetic fields *oppose* the applied field. But the response is usually very weak and so $\mu \sim 1$. This is due primarily to the response of the electron “orbit” in an atom.

Paramagnetic materials

Some materials have a natural net magnetic dipole moment. Such materials give a paramagnetic response, which can arise from dipole moments in an atom, molecules, crystal defects, and conduction electrons. The dipole moments tend to align with the magnetic field but are deflected from complete alignment by their thermal activity. The fields resulting from the partial alignment *adds* to the applied field so $\mu \geq 1$. It can be shown that this effect is also relatively weak so that at room temperature $\mu \approx 1 + 10^{-5}$ [27].

Ferromagnetics and ferrimagnetics

Materials with residual magnetization exist where it is energetically favorable for internal magnetic dipoles to align spontaneously below a certain (Curie) temperature. These *ferromagnetic* and *ferrimagnetic* materials exhibit non-linear and large μ factors $\sim 10^3 - 10^6$. As shown in Fig. 4.5, these magnetic moments align in response to the external field. Note that it is energetically favorable to align in the direction of the field, thus enhancing the magnetic field. In diamagnetic materials, by contrast, the effect of the material is to reduce the field.

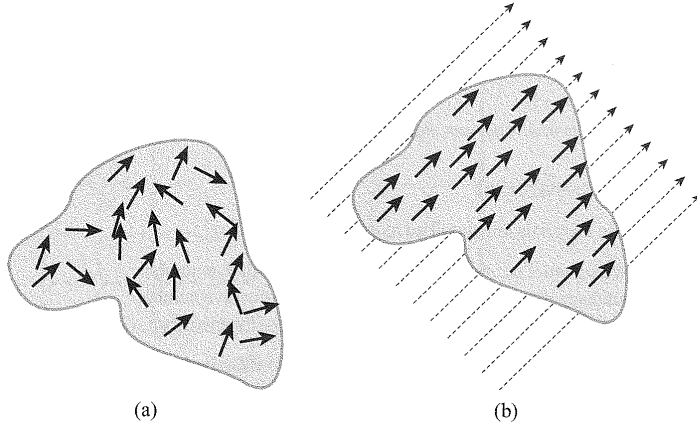


Figure 4.5 For ferromagnetic materials, (a) in the absence of an external field, the magnetic domains point in random directions. (b) The application of the field aligns the magnetic dipoles.

Boundary conditions for a magnetic field

We have now established the following equations for a static magnetic field

$$\nabla \times \mathbf{H} = \mathbf{J} \quad (4.66)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (4.67)$$

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} = I \quad (4.68)$$

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0 \quad (4.69)$$

And for linear materials, we find that $\mathbf{H} = \mu^{-1}\mathbf{B}$

Tangential $\mathbf{H}$

The appropriate boundary conditions follow immediately from our previously established techniques. As shown in Fig. 4.6a, take a small loop intersecting with the boundary and take the limit as the loop becomes tiny

$$\int_C \mathbf{H} \cdot d\mathbf{l} = (H_{t1} - H_{t2})d\ell = 0 \quad (4.70)$$

So the tangential component of H is continuous

$$H_{t1} = H_{t2} \quad (4.71)$$

$$\mu_1^{-1}B_{t1} = \mu_2^{-1}B_{t2} \quad (4.72)$$

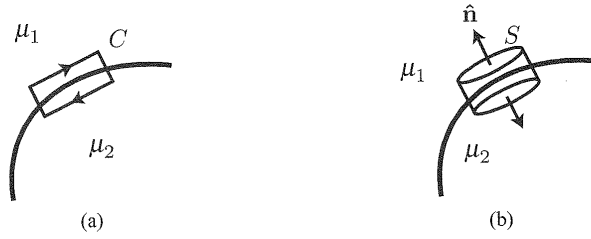


Figure 4.6 The setup for the calculation of the boundary condition for the B field.

Note that B is discontinuous because there is an effective surface current due to the change in permeability. Since B is “real,” it reflects this change. If, in addition, a surface current is flowing in between the regions, then we need to include it in the above calculation.

Normal B

Consider a pillbox cylinder enclosing the boundary between the layers, as shown in Fig. 4.6b. In the limit that the pillbox becomes small, we have

$$\oint \mathbf{B} \cdot d\mathbf{S} = (B_{1n} - B_{2n})dS = 0 \quad (4.73)$$

and thus the normal component of B is continuous

$$B_{1n} = B_{2n} \quad (4.74)$$

Boundary conditions for a conductor

If a material is a very good conductor, then we will show that it can only support current at the surface of the conductor. In fact, for an ideal conductor, the current lies entirely on the surface and it is a true surface current. In such a case the current enclosed by an infinitesimal loop is finite

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = (H_{t1} - H_{t2})d\ell = J_s d\ell \quad (4.75)$$

$$H_{t1} - H_{t2} = J_s \quad (4.76)$$

This can be expressed compactly as

$$\hat{\mathbf{n}} \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_s \quad (4.77)$$

But for a perfect conductor, we will see that $H_2 = 0$, so $H_{t1} = J_s$

$$\hat{\mathbf{n}} \times \mathbf{H}_1 = \mathbf{J}_s \quad (4.78)$$

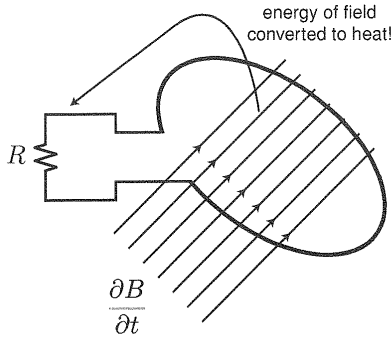


Figure 4.7 A changing magnetic field cuts through a loop and generates a current. The magnetic energy is converted to heat in the load resistor.

4.3 Faraday's big discovery

In electrostatics, we learned that $\oint \mathbf{E} \cdot d\mathbf{l} = 0$. Let's use the analogy between B and D (and E and H). Since $q = Cv$ and $\psi = Li$, and $i = \dot{q} = C\dot{v}$, should we not expect that $\dot{\psi} = L\dot{i} = v$? In fact, this is true! Faraday was able to show this experimentally

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\frac{d\psi}{dt} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{S} \quad (4.79)$$

The electric force is no longer strictly conservative, $\mathbf{E} \neq -\nabla\phi$.

Faraday's Law in differential form

Using Stoke's Theorem

$$\int_C \mathbf{E} \cdot d\mathbf{l} = \int_S \nabla \times \mathbf{E} \cdot d\mathbf{S} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{S} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S} \quad (4.80)$$

Since this is true for any arbitrary curve C , this implies that

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (4.81)$$

Faraday's Law is true for any region of space, including free space. In particular, if C is bounded by an actual loop of wire, then the flux cutting this loop will induce a voltage around the loop. A simple experimental setup is shown in Fig. 4.7 where the magnetic energy is converted to heat.

Example 8

Transformers

A transformer is constructed as shown in Fig. 4.8. In a transformer, by definition the flux in the "primary" side is given by $\psi_1 = L_1 I_1$. Likewise, the flux crossing the

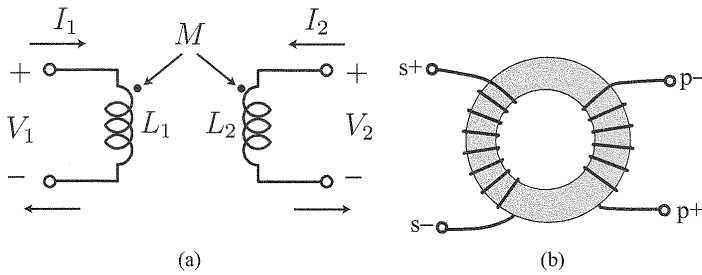


Figure 4.8 (a) A transformer circuit element is constructed using (b) two windings around a common magnetic core.

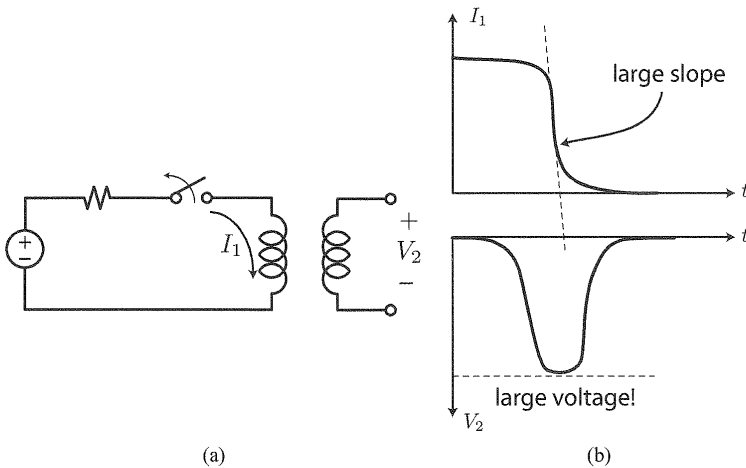


Figure 4.9 (a) A simple spark-gap circuit works by interrupting the current in the primary, thus generating a large magnetic flux change and (b) a large voltage is induced in the secondary terminals.

"secondary" side is given by $\psi_2 = M_{21}I_1 = M_{12}I_1 = MI_1$ (assuming $I_2 = 0$). Thus if the current in the primary changes, a voltage is induced in the secondary

$$V_2 = \dot{\psi}_2 = M\dot{I}_1 \quad (4.82)$$

Generating sparks!

Since the voltage at the secondary side is proportional to the rate of change of current in loop 1, we can generate very large voltages at the secondary side by interrupting the current with a switch. This is shown schematically in Fig. 4.9. This is in fact a simplified version of how spark plugs are fired in an automobile engine.

The return of the vector potential

Since $\nabla \cdot \mathbf{B} \equiv 0$, we can write $\mathbf{B} = \nabla \times \mathbf{A}$. Thus

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\frac{\partial \nabla \times \mathbf{A}}{\partial t} = -\frac{\partial \mathbf{A}}{\partial t} \quad (4.83)$$

If we group terms, we have

$$\left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = 0 \quad (4.84)$$

So, as we saw in electrostatics, we can likewise write

$$\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla \phi \quad (4.85)$$

We choose a negative sign for ϕ to be consistent with electrostatics. Since if $\frac{\partial}{\partial t} = 0$, this equation breaks down to the electrostatic case and then we identify ϕ as the scalar potential. This gives us some insight into the electromagnetic response as

$$\mathbf{E} = -\nabla \phi - \frac{\partial \mathbf{A}}{\partial t} \quad (4.86)$$

$$\mathbf{E} = \underbrace{-\nabla \phi}_{\text{electric response}} - \underbrace{\frac{\partial \mathbf{A}}{\partial t}}_{\text{magnetic response}} \quad (4.87)$$

Is the vector potential real?

We can now re-derive Faraday's Law as follows

$$V = \oint_C \mathbf{E} \cdot d\mathbf{l} = -\oint_C \nabla \phi \cdot d\mathbf{l} - \frac{\partial}{\partial t} \oint_C \mathbf{A} \cdot d\mathbf{l} \quad (4.88)$$

The line integral involving $\nabla \phi$ is zero by definition, so we have the induced emf equal to the line integral of $\mathbf{A}$ around the loop in question

$$V = -\frac{\partial}{\partial t} \oint_C \mathbf{A} \cdot d\mathbf{l} \quad (4.89)$$

We also found that equivalently

$$V = -\frac{\partial}{\partial t} \int_S \mathbf{B} \cdot d\mathbf{S} \quad (4.90)$$

Eq. (4.89) is somewhat more satisfying than Faraday's Law in terms of the flux. Although it is mathematically equivalent, it explicitly shows us the shape of the loop's role in determining the induced flux. The flux equation, though, depends on a surface bounding the loop, and *any* surface will do (this is discussed in more detail in the next chapter). Sometimes it is even difficult to imagine the shape of such a surface (e.g. a coil).

Consider the magnetic coupling between a solenoid and a large loop surrounding the solenoid, as shown in Fig. 4.10. For an ideal solenoid, $\mathbf{B} = 0$ outside of the cylinder. Certainly we can assume that $\mathbf{B} \approx 0$ outside of this region.

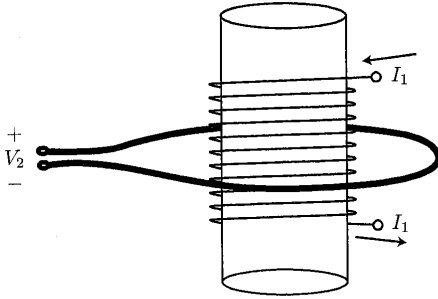


Figure 4.10 An ideal solenoid has zero magnetic field outside of the loop.

Then the voltage induced into the outer loop only depends on the constant flux generated within the center section coincident with the solenoid. What is disturbing is that, even though $\mathbf{B} = 0$ along the loop, there is a force pushing electrons inside the metal. The force would therefore appear to arise *not* from a magnetic field since $\mathbf{B} = 0$. The viewpoint with vector potential, though, does not pose any problems since $\mathbf{A} \neq 0$ outside of the loop. Therefore, when we integrate $\mathbf{A}$ outside of the loop, there is a non-zero result.

4.4 Maxwell's displacement current

Let's summarize what we have learned thus far. There are no magnetic charges, so

$$\nabla \cdot \mathbf{B} = 0 \quad (4.91)$$

and electric fields diverge on physical charge

$$\nabla \cdot \mathbf{D} = \rho \quad (4.92)$$

Faraday's Law tells us that

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (4.93)$$

and Ampère's Law relates magnetic fields to currents by

$$\nabla \times \mathbf{H} = \mathbf{J} \quad (4.94)$$

Are these equations complete and self consistent? In other words, do they over specify the problem or are some equations still missing? Furthermore, are they self consistent? Mathematics tells us that $\nabla \cdot (\nabla \times \mathbf{H}) = 0$, which implies that

$$\nabla \cdot \mathbf{J} = 0 \quad (4.95)$$

But this can only hold for steady fields. In general, by conservation of charge we know that

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} \quad (4.96)$$

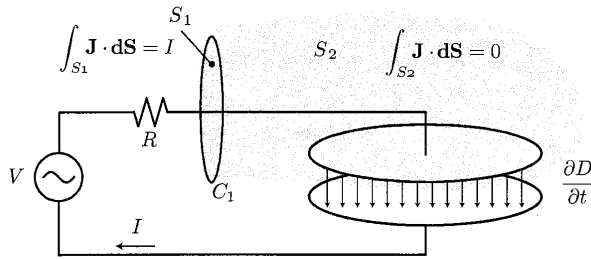


Figure 4.11 The magnetic field can be computed by surface S_1 or S_2 . The surface S_1 cuts through a wire, and thus by Ampère's Law generates a magnetic field. But the surface S_2 does not cut through any DC current.

In other words, we have to add something to the right-hand side of Ampère's equation to make it self consistent! Maxwell was the first to make this observation. Since $\nabla \cdot \mathbf{D} = \rho$, it is natural to add a *displacement current* to the Ampère's equation

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (4.97)$$

This now makes our equation self consistent since

$$\nabla \cdot \nabla \times \mathbf{H} = 0 = \nabla \cdot \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} = \nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} \quad (4.98)$$

Magnetic field of a capacitor

Now we can resolve a contradiction in Ampère's equation. If we consider the magnetic field of the circuit shown in Fig. 4.11, we know that there is a magnetic field around loop C_1 since current cuts through surface S_1 . But Ampère's Law says that *any* surface bounded by C_1 can be used to calculate the magnetic field. If we use surface S_2 , then the current cutting through this surface is zero, which would yield a zero magnetic field!

Displacement current of a capacitor

The answer to this contradiction is displacement current. If current is flowing in this circuit, then the electric field between the capacitor plates must be changing. Thus $\frac{\partial \mathbf{D}}{\partial t} \neq 0$. So the displacement current cutting surface S_2 must be the same as the conductive current cutting through surface S_1

$$\int_{S_1} \mathbf{J}_c \cdot d\mathbf{S} = \int_{S_2} \frac{\partial \mathbf{D}}{\partial t} \cdot d\mathbf{S} \quad (4.99)$$

Maxwell's equations

We have now studied the complete set of Maxwell's equations. In Integral form

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{S} \quad (4.100)$$

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \frac{d}{dt} \int_S \mathbf{D} \cdot d\mathbf{S} + \int_S \mathbf{J} \cdot d\mathbf{S} \quad (4.101)$$

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = \int_V \rho dV \quad (4.102)$$

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0 \quad (4.103)$$

The fields are related by the following material parameters

$$\mathbf{B} = \mu \mathbf{H} \quad (4.104)$$

$$\mathbf{D} = \epsilon \mathbf{E} \quad (4.105)$$

$$\mathbf{J} = \sigma \mathbf{E} \quad (4.106)$$

For most materials we assume that these are scalar relations. In differential form

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (4.107)$$

$$\nabla \times \mathbf{B} = \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} + \mu_0 \left(\mathbf{J} + \frac{\partial \mathbf{P}}{\partial t} + \nabla \times \mathbf{M} \right) \quad (4.108)$$

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0} (\rho - \nabla \cdot \mathbf{P}) \quad (4.109)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (4.110)$$

Source-free regions

In source-free regions $\rho = 0$ and $\mathbf{J} = 0$. Assume the material is uniform (no bound charges or currents)

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (4.111)$$

$$\nabla \times \mathbf{B} = \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} \quad (4.112)$$

$$\nabla \cdot \mathbf{E} = 0 \quad (4.113)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (4.114)$$

As shown in Fig. 4.12, we can intuitively see that $\frac{\partial \mathbf{E}}{\partial t} \rightarrow \frac{\partial \mathbf{B}}{\partial t} \rightarrow \frac{\partial \mathbf{E}}{\partial t} \rightarrow \dots$, in other words wave motion is possible.

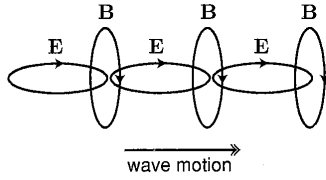


Figure 4.12 The propagation of an electromagnetic wave in a source-free region occurs as the solenoidal electric field generates a magnetic field, and a magnetic field in turn generates an electric field.

Time-harmonic Maxwell's equation

Under time-harmonic conditions (many important practical cases are time harmonic, or nearly so, or else Fourier analysis can handle non-harmonic cases)

$$\nabla \times \mathbf{E} = -j\omega\mathbf{B} \quad (4.115)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + j\omega\mathbf{D} \quad (4.116)$$

$$\nabla \cdot \mathbf{E} = \rho/\epsilon \quad (4.117)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (4.118)$$

These equations are not all independent. Take the divergence of the curl, for instance

$$\nabla \cdot (\nabla \times \mathbf{E}) \equiv 0 = -j\omega \nabla \cdot \mathbf{B} \quad (4.119)$$

In other words, the non-existence of magnetic charge is built in to our curl equation. If magnetic charge is ever observed, we would have to modify our equations. This is analogous to the displacement current that Maxwell introduced to make the curl of $\mathbf{H}$ equation self consistent. Working with (4.116)

$$\nabla \cdot (\nabla \times \mathbf{H}) \equiv 0 = \nabla \cdot \mathbf{J} + j\omega \nabla \cdot \mathbf{D} \quad (4.120)$$

but

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} = -j\omega\rho \quad (4.121)$$

This implies that $\nabla \cdot \mathbf{D} = \rho$, so Gauss' Law is built-in to our curl equations as well.

Tangential boundary conditions

The boundary conditions on the E-field at the interface of two media is

$$\hat{\mathbf{n}} \times (\mathbf{E}_1 - \mathbf{E}_2) = 0 \quad (4.122)$$

Or equivalently, $E_{1t} = E_{2t}$. If magnetic charges are ever found, then this condition will have to include the possibility of a surface magnetic current. The boundary conditions on $\mathbf{H}$ are similar

$$\hat{\mathbf{n}} \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_s \quad (4.123)$$

For the interface of a perfect conductor, for example, a surface current flows so that ($H_2 = 0$)

$$\hat{\mathbf{n}} \times \mathbf{H}_1 = \mathbf{J}_s \quad (4.124)$$

Boundary conditions for current

Applying the “pillbox” argument to the divergence of current

$$\int_V (\nabla \cdot \mathbf{J}) dV = \oint_S \mathbf{J} \cdot d\mathbf{S} = - \int_V \frac{\partial \rho}{\partial t} dV \quad (4.125)$$

in the limit

$$J_{1n} - J_{2n} = - \frac{\partial \rho_s}{\partial t} \quad (4.126)$$

where ρ_s is the surface current. In the static case

$$J_{1n} = J_{2n} \quad (4.127)$$

implies that $\sigma_1 E_1 = \sigma_2 E_2$. This implies that $\rho_s \neq 0$ since $\epsilon_1 E_1 \neq \epsilon_2 E_2$ (unless the ratios of σ match the ratio of ϵ perfectly!)

4.5 References

This chapter draws from many sources, particularly Feynman [18], Cheng [6], and Inan and Inan [62].

5 Inductance

5.1 Introduction

Inductance is an intrinsic part of every physical circuit in which currents flow. Even the simplest circuit shown in Fig. 5.1a cannot be fully understood unless the inductance is taken into account. Even though no inductor is drawn in the figure explicitly, there is a net electromagnetic force

$$V_L = L \frac{dI}{dt} \quad (5.1)$$

induced around the loop that tends to oppose any change in the current magnitude. The inductance L is a property of the loop related to the physical geometry of the circuit. We can represent this induced voltage by an element that has the voltage–current relationship embodied in Eq. (5.1), as shown schematically in Fig. 5.1b. In this case, the inductance is a property of the loop and it does not stem from a separate coil (as suggested by the inductor’s symbol) introduced into the circuit. While we can calculate the resistance of the loop by breaking the loop into three separate conductors, as suggested by $R_1 - R_3$, this cannot be done easily with the inductance because the currents in the conductors interact magnetically.

But is this intrinsic inductance really important? Using techniques of this chapter we can determine the magnitude of L for a given circuit. From the application at hand, we can estimate dI/dt or ωL in a steady-state periodic circuit and from Eq. (5.1) we can estimate the magnitude of the induced voltage. If L is sufficiently small, then we can ignore the magnetic effects of the circuit. Clearly then, for circuits operating at higher frequencies or at high switching rates dI/dt , we cannot ignore the inductive behavior. But even at lower frequencies, if the circuit is physically large then the L term is significant and the induced voltages are important.

It is interesting to note that in this circuit there is no one point where this voltage is induced but rather it is induced around the entire loop. In fact, voltages are induced in the entire space and not just along the circuit path. This gives rise to *eddy currents*, or magnetically induced currents that flow in other conducting bodies. Furthermore, the voltage is not induced uniformly along the path of the circuit, but in a distributed fashion where the induction is stronger at points where more magnetic energy is stored.

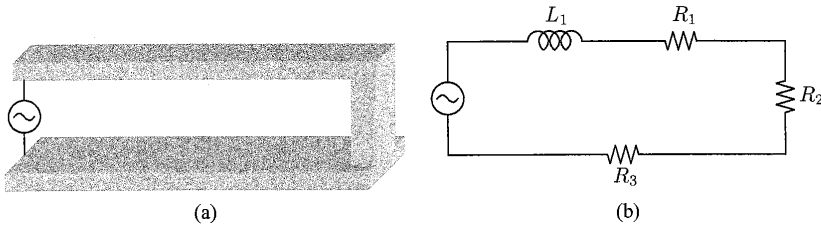


Figure 5.1 (a) A simple circuit with inductance. (b) Schematic representation of circuit.

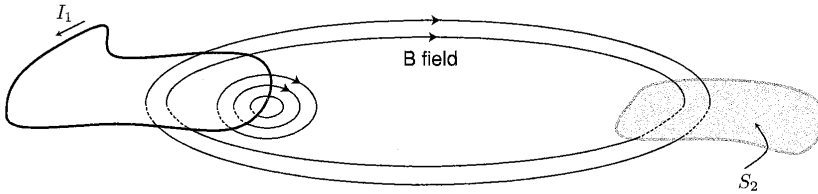


Figure 5.2 The flux linkage between two current loops.

Is the distributed nature of inductance important? Consider the input impedance of the above circuit seen from the terminals of the voltage source. Isn't it simply $j\omega L + R$? Surprisingly, no! In later chapters we will compute the input impedance of this distributed circuit and find the conditions under which the value deviates from the lumped circuit approximation.

5.2 Inductance

Magnetic flux

Magnetic flux plays an important role in many EM problems (in analogy with electric charge)

$$\Psi = \int_S \mathbf{B} \cdot d\mathbf{S} \quad (5.2)$$

Due to the absence of magnetic charge

$$\Psi = \oint_S \mathbf{B} \cdot d\mathbf{S} \equiv 0 \quad (5.3)$$

but net flux can certainly cross an open surface. The magnetic flux crossing a loop is shown in Fig. 5.2 by the integral of B crossing the shaded surface. Magnetic flux is independent of the surface but only depends on the curve bounding the surface. This is easy to show

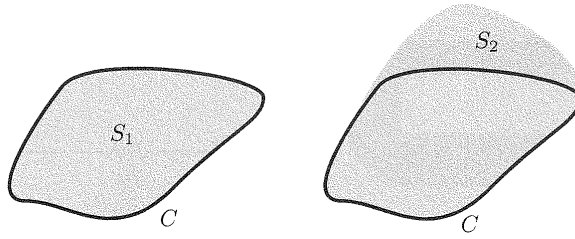


Figure 5.3 Any open surface bounded by the circuit can be used to calculate the magnetic flux.

since

$$\Psi = \int_S \mathbf{B} \cdot d\mathbf{S} = \int_S \nabla \times \mathbf{A} \cdot d\mathbf{S} = \oint_C \mathbf{A} \cdot d\mathbf{l} \quad (5.4)$$

$$\Psi = \int_{S_1} \mathbf{B} \cdot d\mathbf{S} = \int_{S_2} \mathbf{B} \cdot d\mathbf{S} \quad (5.5)$$

So in Fig. 5.2, we can select either S_1 or S_2 to calculate the flux.

Flux linkage

In Fig. 5.3, consider the flux-crossing surface S_2 due to a current I_1 flowing in loop S_1

$$\Psi_{21} = \int_{S_2} \mathbf{B}_1 \cdot d\mathbf{S} \quad (5.6)$$

Likewise, the “self-flux” of a loop is defined by the flux crossing the surface of a path when a current is flowing in the path

$$\Psi_{11} = \int_{S_1} \mathbf{B}_1 \cdot d\mathbf{S} \quad (5.7)$$

The flux is linearly proportional to the current and otherwise only a function of the geometry of the path. To see this, let us calculate Ψ_{21} for filamental loops

$$\Psi_{21} = \oint_{C_2} \mathbf{A}_1 \cdot d\mathbf{l}_2 \quad (5.8)$$

but

$$\mathbf{A}_1 = \frac{1}{4\pi\mu_0^{-1}} \oint_{C_1} \frac{I_1 d\mathbf{l}_1}{R - R_1} \quad (5.9)$$

Substituting, we have a double integral

$$\Psi_{21} = \frac{I_1}{4\pi\mu_0^{-1}} \oint_{C_2} \left(\oint_{C_1} \frac{d\mathbf{l}_1}{R - R_1} \right) \cdot d\mathbf{l}_2 \quad (5.10)$$

The geometry of this double integral is shown in Fig. 5.4. We thus have a simple formula that only involves the magnitude of the current and the average distance between every two

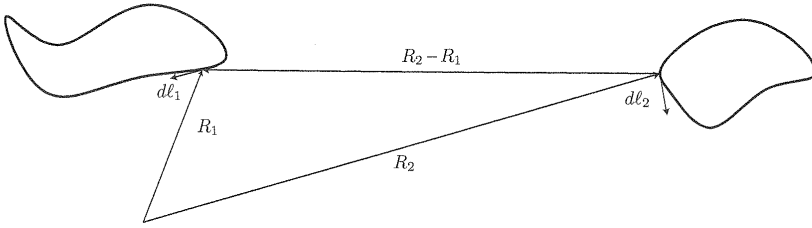


Figure 5.4 The distance between two arbitrary points in the loop.

points on the loops

$$\Psi_{21} = I_1 \times \frac{1}{4\pi\mu_0^{-1}} \oint_{C_2} \oint_{C_1} \frac{\mathbf{dl}_1 \cdot \mathbf{dl}_2}{R_2 - R_1} \quad (5.11)$$

Mutual and self inductance

Since the flux is proportional to the current by a geometric factor, we may write

$$\Psi_{21} = M_{21} I_1 \quad (5.12)$$

We call the factor M_{21} the mutual inductance

$$M_{21} = \frac{\Psi_{21}}{I_1} = \frac{1}{4\pi\mu_0^{-1}} \oint_{C_2} \oint_{C_1} \frac{\mathbf{dl}_1 \cdot \mathbf{dl}_2}{R_2 - R_1} \quad (5.13)$$

The above expression is known as Neumann's equation. It illustrates compactly the geometric nature of mutual inductance. In the above computation R takes on the values of the distance between every pair of points on the loops. In other words, mutual inductance is the weighted average value of $1/R$ between every pair of points on the loops weighted by $\cos(\theta)$, where θ is the local angle between the points in the loop. Thus, points lying on segments mutually perpendicular do not contribute to the mutual inductance.

An interesting observation is that mutual coupling is reciprocal. In other words

$$M_{12} = \frac{\mu}{4\pi} \oint_{C_1} \oint_{C_2} \frac{1}{R} \mathbf{dl}_2 \cdot \mathbf{dl}_1 = M_{21} \quad (5.14)$$

since the weighted “averaging” may be done in any order. This is a profound result that has many implications in circuit theory. Consider the example shown in Fig. 5.5 where a small loop couples to a larger loop. Even though the magnetic flux of loop B is small, loop A “captures” a large portion of the flux. Similarly, even though loop A has a large magnetic flux, only a small portion of it is captured by loop B and thus $M_{AB} = M_{BA}$.

The units of M are H since $[\mu] = \text{H/m}$. The “self-flux” mutual inductance is simply called the self inductance and denoted by $L_1 = M_{11}$.

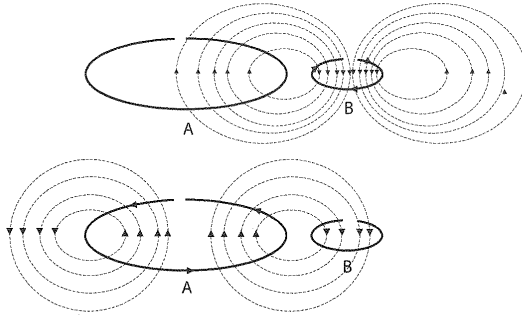


Figure 5.5 Illustration of reciprocity through the *forward* and *reverse* coupling between two vastly different loops.

System of mutual inductance equations

If we generalize to a system of current loops we have a system of equations

$$\begin{aligned}\Psi_1 &= L_1 I_1 + M_{12} I_2 + \dots M_{1N} I_N \\ &\vdots \\ \Psi_N &= M_{N1} I_1 + M_{N2} I_2 + \dots L_N I_N\end{aligned}$$

Or in matrix form $\psi = M\mathbf{i}$, where M is the inductance matrix. This equation resembles $\mathbf{q} = C\mathbf{v}$, where C is the capacitance matrix.

Example 9

Coaxial conductor

In transmission line problems, we need to compute inductance/unit length. Consider the shaded area in Fig. 5.6, from $r = a$ to $r = b$. The magnetic field in the region between conductors is easily computed

$$\oint \mathbf{B} \cdot d\mathbf{l} = B_\phi 2\pi r = \mu_0 I \quad (5.15)$$

The external flux (excluding the volume of the ideal conductors) is given by

$$\psi' = \int_a^b B_\phi dr = \frac{\mu_0 I}{2\pi} \int_a^b \frac{dr}{r} = \frac{\mu_0 I}{2\pi} \ln\left(\frac{b}{a}\right) \quad (5.16)$$

The inductance per unit length is therefore

$$L' = \frac{\mu_0}{2\pi} \ln\left(\frac{b}{a}\right) \text{ [H/m]} \quad (5.17)$$

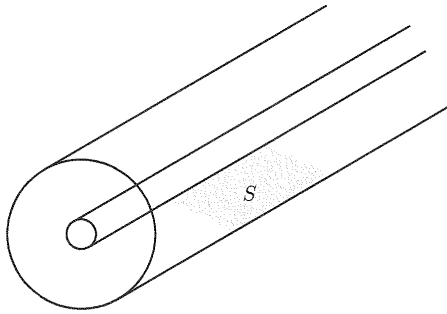


Figure 5.6 The partial inductance of a coaxial conductor is computed by calculating the flux in the shaded area.

Recall that the product of inductance and capacitance per unit length for a transmission line is a constant (see Ch. 9)

$$L'C' = \frac{1}{c^2} \quad (5.18)$$

where c is the speed of light in the medium. Thus we have calculated the capacitance per unit length without any extra work.

5.3 Magnetic energy and inductance

Energy for a system of current loops

We would like to calculate the magnetic energy associated with a system of current loops, shown in Fig. 5.7. In the electrostatic case, we assembled our charge distribution one point charge at a time and used electric potential to calculate the energy. This can be done for the magnetostatic case but there are some complications.

Energy for two loops

As we move in our second loop with current I_2 , we would be cutting across flux from loop 1, as shown in Fig. 5.8a, and therefore an induced voltage around loop 2 would change the current. When we bring the loop to rest, the induced voltage would drop to zero. To maintain a constant current, therefore, we would have to supply a voltage source in series to cancel the induced voltage. The work done by this voltage source represents the magnetostatic energy in the system.

A simpler approach is to bring in the two loops with zero current and then increase the current in each loop one at a time, shown in Fig. 5.8b. First, let us increase the current in

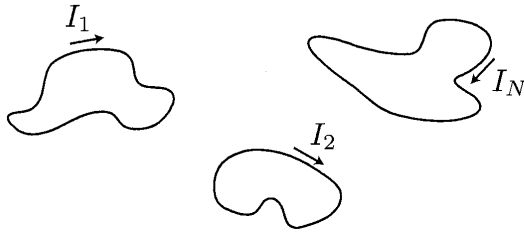


Figure 5.7 A system of current loops store magnetic energy.

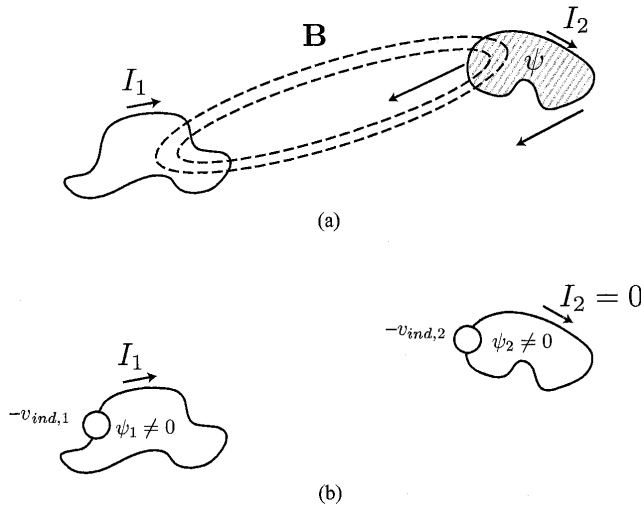


Figure 5.8 (a) When moving loop 2 into our system, the rate of change of flux in loop 2 generates a current in loop 2, which changes the field configuration. (b) Instead, we bring in loops 1 and 2 without any current, and turn on the current in each loop one at a time.

loop 1 from zero to I_1 in some time t_1 . Note that at any instant of time, a voltage is induced around loop number 1 due to its changing flux

$$v_{ind,1} = -\frac{d\psi}{dt} = -L_1 \frac{di_1}{dt} \quad (5.19)$$

where i_1 represents the instantaneous current.

Note that this induced voltage will tend to decrease the current in loop 1. This is a statement of Lenz's Law. In other words, the induced voltage in loop 1 tends to create a magnetic field to oppose the field of the original current! To keep the current constant in loop 1, we must connect a voltage source to cancel the induced voltage. The work done by

this voltage source is given by

$$w_1 = \int_0^{t_1} p(\tau) d\tau = - \int_0^{t_1} v_{ind,1}(\tau) d\tau \quad (5.20)$$

where $p(t) = -v_{ind,1}i_1(t) = L_1 i_1 \frac{di_1}{dt}$. The net work done by the source is simply

$$w_1 = L_1 \int_0^{t_1} i_1 \frac{di_1}{d\tau} d\tau = L_1 \int_0^{I_1} i_1 di_1 = \frac{1}{2} L_1 I_1^2 \quad (5.21)$$

Note that to keep the current in loop 2 equal to zero, we must also provide a voltage source to counter the induced voltage

$$v_{ind,2}(t) = -M_{21} \frac{di_1}{dt} \quad (5.22)$$

This voltage source does not do any work since $i_2(t) = 0$ during this time.

By the same argument, if we increase the current in loop 2 from 0 to I_2 in time t_2 , we need to do work equal to $\frac{1}{2} L_2 I_2^2$. But is that all? No, since to keep the current in loop 1 constant at I_1 we must connect a voltage source to cancel the induced voltage

$$-v_{ind,1} = \frac{d\psi_1}{dt} = M_{12} \frac{di_2}{dt} \quad (5.23)$$

The additional work done is therefore

$$w'_1 = \int_0^{t_2} M_{12} \frac{di_2}{d\tau} I_1 d\tau = M_{12} I_1 I_2 \quad (5.24)$$

The total work to bring the current in loop 1 and loop 2 to I_1 and I_2 is therefore

$$W = \frac{1}{2} L_1 I_1^2 + \frac{1}{2} L_2 I_2^2 + M_{12} I_1 I_2 \quad (5.25)$$

But the energy should not depend on the order we turn on the currents. Thus we can immediately conclude that $M_{12} = M_{21}$. We already saw this when we derived an expression for M_{12} using the Neumann equation.

Generalize to N loops

We can now pretty easily generalize our argument for two loops to N loops

$$W = \frac{1}{2} \sum_i L_i I_i^2 + \sum_{i>j} \sum M_{ij} I_i I_j \quad (5.26)$$

The first term represents the “self” energy for each loop and the second term represents the interaction terms.

$$W = \frac{1}{2} \sum_{i=j} \sum M_{ij} I_i I_j + \frac{1}{2} \sum_{i \neq j} \sum M_{ij} I_i I_j \quad (5.27)$$

$$W = \frac{1}{2} \sum_i \sum_j M_{ij} I_i I_j \quad (5.28)$$

The factor $1/2$ accounts for the double counting in the double summation. Let's rewrite this equation and combine terms

Energy in terms of vector potential

We derived the mutual inductance between two filamentary loops using Neumann's equation

$$M_{ij} = \frac{\mu_0}{4\pi} \oint_{C_i} \oint_{C_j} \frac{\mathbf{dl}_i \cdot \mathbf{dl}_j}{R_{ij}} \quad (5.29)$$

Let's substitute the above relation into the expression for energy

$$W = \frac{1}{2} \sum_i I_i \left[\sum_j M_{ij} I_j \right] \quad (5.30)$$

$$W = \frac{1}{2} \sum_i I_i \left[\sum_j I_j \frac{\mu_0}{4\pi} \oint_{C_i} \oint_{C_j} \frac{\mathbf{dl}_i \cdot \mathbf{dl}_j}{R_{ij}} \right] \quad (5.31)$$

Let's change the order of integration and summation

$$W = \frac{1}{2} \sum_i I_i \oint_{C_i} \left[\sum_j I_j \frac{\mu_0}{4\pi} \oint_{C_j} \frac{\mathbf{dl}_j}{R_{ij}} \right] \cdot \mathbf{dl}_i \quad (5.32)$$

Each term of the bracketed expression represents the vector potential due to loop j evaluated at a position on loop i . By superposition, the sum represents the total voltage potential due to all loops

$$W = \frac{1}{2} \oint_{C_i} I_i \mathbf{A} \cdot \mathbf{dl}_i \quad (5.33)$$

We derived this for filamental loops. Generalize to an arbitrary current distribution and we have

$$W_m = \frac{1}{2} \int_V \mathbf{J} \cdot \mathbf{A} dV \quad (5.34)$$

Compare this to the expression for electrostatic energy

$$W_e = \frac{1}{2} \int_V \rho \phi dV \quad (5.35)$$

Thus the vector potential $\mathbf{A}$ really does represent the magnetic potential due to a current distribution in an analogous fashion as ϕ represents the electric potential.

Energy in terms of fields

Let's replace $\mathbf{J}$ by Ampère's Law $\nabla \times \mathbf{H} = \mathbf{J}$

$$W_m = \frac{1}{2} \int_V (\nabla \times \mathbf{H}) \cdot \mathbf{A} dV \quad (5.36)$$

Using the identity

$$\nabla \cdot (\mathbf{H} \times \mathbf{A}) = (\nabla \times \mathbf{H}) \cdot \mathbf{A} + \mathbf{H} \cdot (\nabla \times \mathbf{A}) \quad (5.37)$$

$$W_m = \frac{1}{2} \int_V \nabla \cdot (\mathbf{H} \times \mathbf{A}) dV + \frac{1}{2} \int_V (\nabla \times \mathbf{A}) \cdot \mathbf{H} dV \quad (5.38)$$

Apply the Divergence Theorem to the first term to give

$$W_m = \frac{1}{2} \int_S \mathbf{H} \times \mathbf{A} \cdot d\mathbf{S} + \frac{1}{2} \int_V (\nabla \times \mathbf{A}) \cdot \mathbf{H} dV \quad (5.39)$$

We would like to show that the first term is zero. To do this, consider the energy in all of space $V \rightarrow \infty$. Use a large sphere of radius r and take the radius to infinity. We know that if we are sufficiently far from the current loops, the potential and field behave like $A \sim r^{-1}$ and $H \sim r^{-2}$. The surface area of the sphere goes like r^2 . The surface integral, therefore, gets smaller and smaller as the sphere approaches infinity.

The remaining volume integral represents the total magnetic energy of a system of currents

$$W_m = \frac{1}{2} \int_V (\nabla \times \mathbf{A}) \cdot \mathbf{H} dV \quad (5.40)$$

But $\nabla \times \mathbf{A} = \mathbf{B}$

$$W_m = \frac{1}{2} \int_V \mathbf{B} \cdot \mathbf{H} dV \quad (5.41)$$

And the energy density of the field is seen to be

$$w_m = \mathbf{B} \cdot \mathbf{H} \quad (5.42)$$

Recall that $w_e = \mathbf{D} \cdot \mathbf{E}$

Another formula for inductance

The self inductance of a loop is given by

$$L = \frac{1}{I} \int_S \mathbf{B} \cdot d\mathbf{S} \quad (5.43)$$

Since the total magnetic energy for a loop is $\frac{1}{2} L I^2$, we have an alternate expression for the inductance

$$\frac{1}{2} L I^2 = \frac{1}{2} \int_V \mathbf{B} \cdot \mathbf{H} dV \quad (5.44)$$

$$L = \frac{1}{I^2} \int_V \mathbf{B} \cdot \mathbf{H} dV \quad (5.45)$$

This alternative expression is sometimes easier to calculate. We will see this when we compute the *internal* inductance of a round wire.

Self inductance of filamentary loops

We have tacitly assumed that the inductance of a loop is a well-defined quantity. But for a filamentary loop, we can expect trouble. By definition

$$L = \frac{1}{I} \int_S \mathbf{B} \cdot d\mathbf{S} = \frac{1}{I} \oint_C \mathbf{A} \cdot d\mathbf{l} \quad (5.46)$$

$$L = \frac{\mu}{4\pi} \oint_C \oint_C \frac{d\ell' \cdot d\ell}{R} \quad (5.47)$$

This is just Neumann's equation with $C_1 = C_2$. But for a filamental loop, $R = 0$ when both loops traverse the same point. The integral is thus not defined for a filamental loop!

Magnetic energy of a circuit

In the previous section we derived that the total magnetic energy in a circuit is given by

$$E_m = \frac{1}{2} L_1 I_1^2 + \frac{1}{2} L_2 I_2^2 + M I_1 I_2 \quad (5.48)$$

We would like to show that this implies that $M \leq \sqrt{L_1 L_2}$. Let's re-write the above into the following positive definite form

$$E_m = \frac{1}{2} L_1 \left(I_1 + \frac{M}{L_1} I_2 \right)^2 + \frac{1}{2} \left(L_2 - \frac{M^2}{L_1} \right) I_2^2 \quad (5.49)$$

An important observation is that regardless of the current I_1 or I_2 , the magnetic energy is non-negative, so $E_m \geq 0$.

Consider the current $I_2 = -\frac{L_1}{M} I_1$, which cancels the first term in E_m

$$E_m = \frac{1}{2} \left(L_2 - \frac{M^2}{L_1} \right) I_2^2 \geq 0 \quad (5.50)$$

Since $I_2^2 \geq 0$, we have

$$L_2 - \frac{M^2}{L_1} \geq 0 \quad (5.51)$$

Therefore it is now clear that this implies

$$L_1 L_2 \geq M^2 \quad (5.52)$$

Coupling coefficient

Usually we express this inequality as

$$M = k \sqrt{L_1 L_2} \quad (5.53)$$

where k is the *coupling coefficient* between two circuits, with $|k| \leq 1$. If two circuits are perfectly coupled (all flux from circuit 1 crosses circuit 2), $k = 1$ (ideal transformer). Note that $M < 0$ implies that $k < 0$, which is totally reasonable as long as k lies on the unit interval $-1 \leq k \leq 1$. Negative coupling just means that the flux gets inverted before crossing the second circuit. This is easily achieved by winding the circuits with opposite orientation.

5.4 Discussion of inductance

As we have seen in the previous section, there are many equivalent ways to define inductance. All of the following forms are equally important as they shed new insights into the problem. In this section we discuss these definitions in further detail.

Magnetostatics and quasistatics

Magnetostatics is the study of magnetic fields under time-invariant conditions. Clearly, with no time variation, Maxwell's equations take on very simple forms. While these conditions seem oversimplified for practical applications, especially when disallowing any time variation, we shall see later that many seemingly difficult non-static problems admit solutions when seen in this light. For instance, when studying transmission lines, structures where electric and magnetic energy is in constant dynamic exchange, we shall find that the cross-sectional waveforms take on static field configurations. In other problems, the static behavior will dominate up to very high frequencies and thus we can solve such problems using the *quasistatic* assumption. At a given frequency, or equivalently for a given wavelength of operation, we shall show that this is especially true for structures that are small relative to the wavelength. For instance, even for signals varying tens of billions of times per second, the wavelength is of the order of centimeters, an order of magnitude larger than most integrated circuit structures.

Under the magnetostatic conditions, currents flow uniformly in the volume of the problem space. Thus the net current flow into any volume element V is zero. If it were non-zero, then net charge would continuously accumulate in this volume element, violating the steady-state assumption.

The net current flowing into the volume element V is given by integrating the component of current density $\mathbf{J}$ perpendicular to the bounding surface of the volume S . By the Divergence Theorem we have

$$0 = \oint_S \mathbf{J} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{J} dV \quad (5.54)$$

Assuming Ohm's Law to be valid in the conductive regions and further assuming that the material is isotropic implies the current density is proportional and parallel to the local electric field $\mathbf{J} = \sigma \mathbf{E}$. Substituting in the above equation we have

$$0 = \int_V \nabla \cdot \mathbf{J} dV = \int_V \sigma \nabla \cdot \mathbf{E} dV \quad (5.55)$$

where for simplicity we assume that the material conductivity σ is independent of position. But from Maxwell's equations, in particular Gauss' Law, the divergence of the electric field can only arise from charge. We therefore have

$$0 = \int_V \frac{\sigma}{\epsilon} \rho dV \quad (5.56)$$

Again, let us assume that the permittivity does not vary with position. Take an arbitrary small conductive volume ΔV . Since $\sigma \neq 0$, this forces the net charge to equal zero, or $\rho = 0$. Since the volume element we chose was arbitrary, this implies that all conductors in the problem must have a net charge of zero.

But current flow is a result of charge in motion and from the outset we assume that $\mathbf{J} \neq 0$. Is $\rho = 0$ a contradiction? No, because at any given moment we assume that all moving charges are neutralized by charges of opposite sign. These charges can also be in motion, such as holes in a semiconductor, or they can be immobile ionized host atoms. In

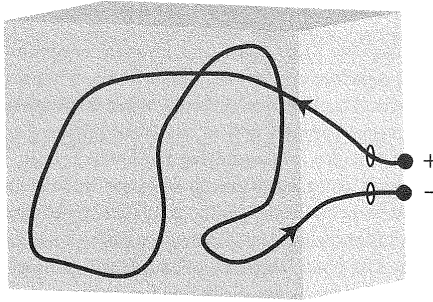


Figure 5.9 A conductive box containing an inductor. The inductor wires are insulated from the box. The inductor may or may not self-intersect.

reality, this is an approximation for finite values of σ since the atoms are in constant thermal agitation causing a net time-varying multi-pole charge moment.

But doesn't $\rho = 0$ imply also that $\mathbf{E} = 0$? No, only that $\nabla \cdot \mathbf{E} = 0$. Since we have assumed steady-state conditions, the electric field has only a curl-free component

$$\mathbf{E} = -\nabla\phi \quad (5.57)$$

The magnetic field does not interact with the electric field since this can only occur under time-varying situations. Maxwell's equations lead to a source-free Poisson's equation

$$\nabla^2\phi = -\frac{\rho}{\epsilon} = 0 \quad (5.58)$$

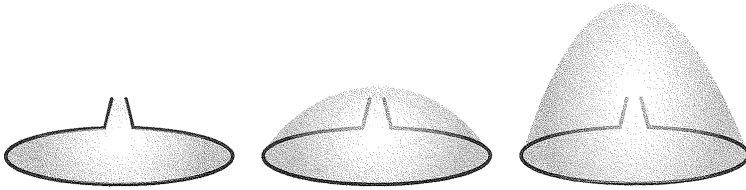
Since there are no sources in the above equation, the scalar potential can be non-zero only if the boundary conditions are non-zero. If we expand the volume space of the problem to infinity where $\phi = 0$, then we do indeed have a contraction. Since in this case we have $\phi \equiv 0$ and thus $\mathbf{E} \equiv 0$, hence no currents can flow. This would also be the case for a grounded closed cavity with walls constructed from perfect conductors. Without any time variation, current cannot flow in the interior of the cavity. Thus, we have to assume that on the boundary of our problem space there are points where $\phi \neq 0$.

As shown in Fig. 5.9, we must break the circuit loop under consideration and apply a field between the terminals of the loop.

Magnetic flux

Inductance defined in terms of magnetic flux is the most common approach, but it requires some subtle observations. As we have seen, the voltage induced in a loop can be obtained by integrating the electric field along the path of the circuit. By Stoke's Theorem we have

$$V_L = -\oint_C \mathbf{E} \cdot d\mathbf{l} = -\int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{S} \quad (5.59)$$

Figure 5.10 Several candidate surfaces bounded by the same curve C .

Note that this is true for any surface S bounded by the circuit path C . From Maxwell's equations we have

$$\nabla \times \mathbf{E} = -\frac{d\mathbf{B}}{dt} \quad (5.60)$$

so that

$$V_L = \frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{S} \quad (5.61)$$

The right-hand side is the magnetic flux through the surface bounded by the current. We defined the magnetic flux ψ by

$$\psi \equiv \int_S \mathbf{B} \cdot d\mathbf{S} \quad (5.62)$$

and it follows that

$$V_L = \frac{d\psi}{dt} \quad (5.63)$$

comparing with Eq. (5.1) we have

$$L \equiv \frac{\psi}{I} \quad (5.64)$$

In computing the flux, which surface do we use in the above integral? In Fig. 5.10 we show several candidate surfaces. Clearly, there are an infinite number of surfaces bounded by the closed curve. Is there a unique surface we can attribute to the path? We might try by limiting the surface to the minimal surface (the so-called “soap film” surface generated by dipping the circuit in a soap bath). But the first step of the above manipulation is justified by Stoke's Theorem which admits *any* surface as long as it is enclosed by the path of the circuit.

The magnetic field $\mathbf{B}$ is computed under magnetostatic conditions described above. For a closed circuit path under magnetostatic conditions, this implies a zero electric field in the volume of perfect conductors. Since the voltage drop around a loop of perfectly conducting material is zero, no voltages can be induced in the loop and so the rate of change flux $\dot{\psi}$ must equal zero. A realistic example of this occurs in a superconductor where the static state is literally frozen into space.

Most practical inductors are constructed from materials of finite non-zero resistivity. We break the loop and introduce an electric field at the terminals. Since the electric field is

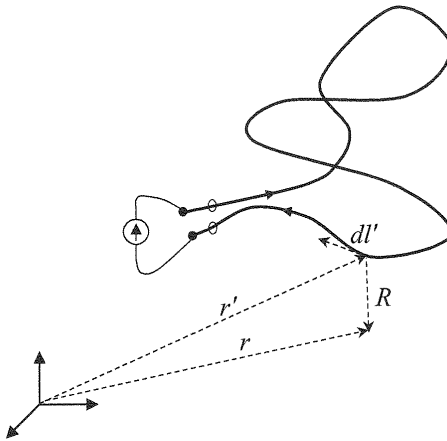


Figure 5.11 Computing the magnetic field using the Biot-Savart equations.

zero or nearly zero along the path of the conductive portion of the loop, the entire induced voltage calculated above appears at the terminals of the circuit.

Under magnetostatic conditions, we can calculate $\mathbf{B}$ from the so-called Biot-Savart Law [52]. This result is easily derived from Ampère's Law in Maxwell's equations

$$\mathbf{B}(r) = \mu \oint_C \frac{I(\mathbf{r}') d\mathbf{l}' \times \mathbf{R}}{4\pi R^3} = \frac{1}{4\pi \mu^{-1}} \oint_C \frac{I(\mathbf{r}') d\mathbf{l}' \times \hat{\mathbf{R}}}{R^2} \quad (5.65)$$

where, as shown in Fig. 5.11, the vector $\mathbf{R}$ is drawn from the “source” point r' along the path of integration on the circuit to the “field” point r . The convention of designating the source point with primed coordinates is an arbitrary but common practice.

The above equation is insightful as it states that the influence of current elements on magnetic field follows a inverse-square law dependence similar to the influence of charge on the electric field.

Now the second subtlety enters in the definition of inductance. In computing the surface integral of Eq. (5.62), where does the boundary of the surface lie? Eq. (5.65) clearly shows that the magnetic field is unbounded as we approach a filamental wire so that the integral is undefined for this case. Therefore, we conclude that the inductance of an infinitesimally thin wire is infinite. Should the boundary be taken as only the inside edge of the wire or should the boundary also include the wire itself?

Internal and external inductance

Let us partition the flux as follows

$$\psi = \psi_i + \psi_o \quad (5.66)$$

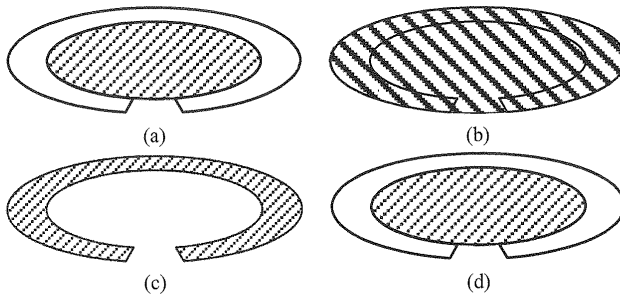


Figure 5.12 (a) The *inner* flux of a circuit. (b) The *outer* flux of a circuit. (c) The *internal* flux of a circuit. (d) The *external* flux of a circuit (same as (a)).

where ψ_i is the inner flux shown in Fig. 5.12a and is computed from the surface integral when the boundary is taken at the internal edge of the loop. The outer flux ψ_o is shown in Fig. 5.12b, where the area of the flux is bounded by the outer edge of the wire forming the loop.

The voltage induced due to $d\psi_i/dt$ is smaller since we are integrating over a smaller area and the flux is non-zero in the volume of the conductors. Comparatively the voltage induced due to $d\psi_o/dt$ is larger. The difference between the voltage induced on the inner edge and outer edge is thus greater the larger the width of the wire. For finite σ we can see then that the current distribution could be non-uniform along the width of the conductor. For good conductors under quasistatic conditions, the current flow is uniform through the width of conductors and the flux is therefore non-zero in the volume of conductors. The voltage induced, therefore, includes both the inner flux ψ_i and the outer flux ψ_o .

The external flux ψ_{ext} is identical with the “inner” flux ψ_i we defined above. The “internal” flux is given by

$$\psi_{int} = \psi_o - \psi_i \quad (5.67)$$

and contains the flux bounded by the inner and outer edges of the conductor. The total inductance is derived by including both terms. This partitioning is convenient once we consider the frequency variation of inductance. As we shall see, at high frequencies the current flow in the volume of conductors decreases (due to induced eddy currents) and current density peaks at the surface of conductors. Then the internal flux ψ_{int} vanishes and the inductance is given by the external flux ψ_{ext} . The external flux is approximately independent of frequency, whereas the internal flux tends to diminish at increasing frequency.

Internal inductance of a round wire

Usually if the wire radius is small relative to the loop area, $L_{int} \ll L_{ext}$. We shall see that at high frequencies, the magnetic field decays rapidly in the volume of conductors and thus the $\psi_{int} \rightarrow 0$ and $L(f \rightarrow \infty) = L_{ext}$. Consider a round wire carrying uniform current. We

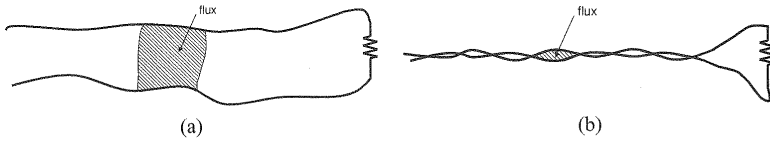


Figure 5.13 (a) Flux of a loop inductor. (b) The magnetic flux of a twisted pair circuit.

can easily derive the magnetic field through Ampère's Law

$$B_{inside} = \frac{\mu_0 I r}{2\pi a^2} \quad (5.68)$$

Using this expression, we can find the internal inductance

$$W_m = \frac{1}{2} \int_V \mathbf{B} \cdot \mathbf{H} dV = \frac{1}{2} \int_{V_{inside}} \mathbf{B} \cdot \mathbf{H} dV + \frac{1}{2} \int_{V_{outside}} \mathbf{B} \cdot \mathbf{H} dV \quad (5.69)$$

$$W_m = \frac{1}{2} L_{int} I^2 + \frac{1}{2} L_{ext} I^2 \quad (5.70)$$

The “inside” term is easily evaluated

$$W_{m,int} = \frac{1}{2} \frac{\mu_0 I^2}{(2\pi a^2)^2} \int_0^a r^2 2\pi r dr = \frac{1}{2} \frac{\mu_0 I^2}{8\pi} \quad (5.71)$$

The internal inductance per unit length is thus

$$L_{int} = \frac{\mu_0}{8\pi} \quad (5.72)$$

Numerically, this is 50pH/mm, a pretty small inductance. Recall that this is only the inductance due to energy stored inside of the wires. The external inductance is likely to be much larger.

Inductance for a general structure

We are now in a position to comment on the inductance of an arbitrary circuit. For instance, to increase the inductance of a circuit we should increase the flux of the circuit. Since the current is divided out of flux, it does not come into play for linear materials where the magnetic permeability μ is independent of current. Hence, the inductance is a purely geometric property of the circuit and we may speak of the inductance in the same way that we speak of other geometric properties, such as the perimeter or surface area of the circuit.

Since the inductance is an integral over the surface of the circuit, it seems as if a circuit that spans a larger area should have higher inductance. For instance, compare the flux area of the two loops shown in Fig. 5.13. In part (a), we show a circuit that spans a large area since the “return current” flows in a wire physically far from the incoming current. In Fig. 5.13b, we keep the return path in close proximity to the incoming current by twisting the wires together. In the limit that the wires are just touching (with some insulation in between to prevent a short), the inductance is driven to arbitrarily low values.

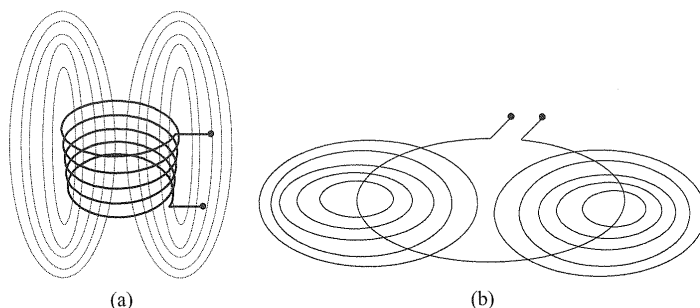


Figure 5.14 (a) A simple loop inductor with area equal to (b) a coil inductor.

Twisting the wires together has another benefit, besides low inductance. The magnetic pickup from a twisted pair is much smaller than a two-wire line with the equivalent inductance. This is because any flux linking with the two-wire line generates an emf, whereas in a twisted pair, for a uniform incoming B field, the flux tends to cancel out since the wire is twisted. In one section of the wire, the flux pickup is positive, but in the next it is negative. In this way, incoming magnetic fields generate a net flux of nearly zero on the twisted pair.

Let us compare two circuits with equal surface area but very different values of inductance. In Fig. 5.14a we see a traditional coil inductor. To help visualize the surface of integration, imagine the portion in the coil as a spiraling ramp resembling a multi-level parking lot structure (Fig. 5.15). For comparison we have also shown a large single loop inductor designed so that the area of the loop is equal to the surface area of the coil. The coil has larger inductance, though, since the magnetic field at any given point inside the coil has additive contributions from several adjacent “loops.” Since the surface area of the single loop is equal to the surface area of the coil, the flux will be larger if the average magnetic field is larger in the coil. Intuitively, it seems like this is the case due to the additive nature of the field from one loop to adjacent loops.

Mutual inductance

The vector nature of the magnetic field leads to an interesting property of magnetic coupling. If the magnetic flux generated by one loop impinges perpendicularly on another loop, it does not generate an induced voltage and hence the mutual coupling $M_{ij} = 0$ for such a case. This principle can be used to isolate structures in a circuit where coupling is undesired.

The net results of mutual inductance can lead to many interesting effects. For instance, loops wound in the opposite orientation add negatively and thus tend to cancel the flux and lower the induced voltage. This is illustrated in Fig. 5.16, where we use the right-hand convention to orient the magnetic fields. Clearly, loop B tends to cancel the self flux of loop A, whereas loop C adds to the flux of loop A.

A nice application of this concept is the ground-fault interrupt circuit (GFI) shown in Fig. 5.17. This circuit is designed to detect a *ground fault*, a condition where an electric

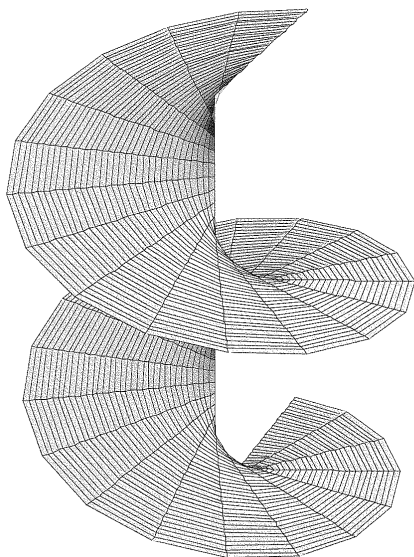


Figure 5.15 A surface bounded by the helix coil inductor.

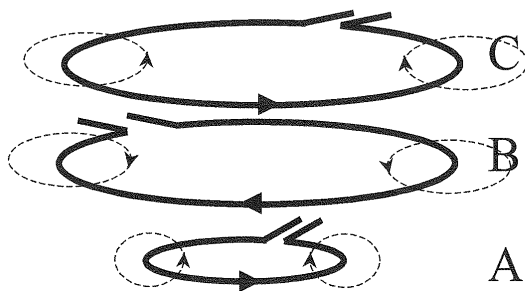


Figure 5.16 Three mutually coupled loops of current. The flux of loop C adds to the flux of loop A, whereas the flux of loop B abstracts from A.

appliance faults by shorting a signal line to ground. Under normal operation all the current that enters the appliance should return to ground via the ground line. If there is any imbalance between the signal line current and ground current, this signifies that a parasitic path to ground has been established, and possibly via conduction current through an unfortunate victim (say a person grounded accidentally touches the high voltage signal line). Under normal operation the voltage on the GFI loop is zero since the flux of the incoming

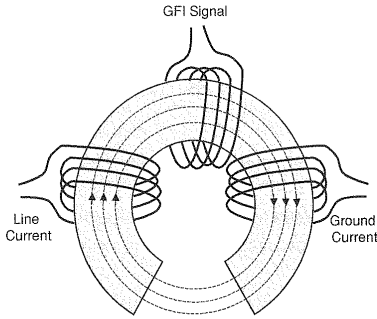


Figure 5.17 Schematic representation of a ground-fault interrupt (GFI) circuit.

current is canceled by the ground current. If any imbalance occurs, a net voltage will be induced at the GFI node and this condition can be detected to quickly shut off power to the appliance.

Magnetic energy perspective

From electromagnetic theory we know that the power stored due to an electromagnetic field is given by [52]

$$P = \frac{1}{2} \frac{d}{dt} \int_V \left(\frac{1}{\mu} |\mathbf{B}|^2 + \epsilon |\mathbf{E}|^2 + \mathbf{E} \cdot \mathbf{J} \right) dV \quad (5.73)$$

where the integral is over the entire volume space of the problem where the fields are non-zero. In a good “inductor,” the magnetic energy dominates over the electrical energy, or $\mu^{-1} |\mathbf{B}|^2 \gg \epsilon |\mathbf{E}|^2$. Thus, the above parenthetical expression reduces to the total stored magnetic energy

$$E_L = \frac{1}{2} \int_V \frac{1}{\mu} |\mathbf{B}|^2 dV \quad (5.74)$$

Using equation E_L we derived an alternate expression for inductance

$$L = \frac{1}{I^2} \int_V \frac{1}{\mu} |\mathbf{B}|^2 dV \quad (5.75)$$

This viewpoint emphasizes the fact that the voltages induced across an inductor terminal or around a circuit loop are due to the magnetic stored energy. Thus an induced voltage can only sustain itself as long as $\mathbf{B}$ is non-zero. In a time periodic circuit the magnetic energy is continuously “pumped” as the induced voltage changes sign. When $dI/dt > 0$ energy is supplied to the inductor. Alternatively, when $dI/dt < 0$ energy returns to the rest of the circuit. Naturally, an inductor constructed from materials with zero resistance must conserve energy.

The inductance computed from the definition of flux must equal the value computed from the energy of the field. This implies that

$$\frac{1}{l} \int_V \frac{1}{\mu} |\mathbf{B}|^2 dV = \int_S \mathbf{B} \cdot d\mathbf{S} \quad (5.76)$$

Example 10

In this example we compute the partial inductance of a two wire transmission line. Assume each wire is round and of radius a and the origins of the wires are separated by a distance d .

From previous examples we already have the necessary tools to compute the external inductance. Since the magnetic field outside of a round wire is equivalent to the field of a filament placed at the origin of the wire, we can reuse the results from the previous example for the magnetic field of two filaments. For the external inductance, we integrate this expression between the wires (excluding the conductors)

$$L = \int_a^{d-a} B_y(y=0) dx \quad (5.77)$$

Recall that $B_y(0)$ is given by

$$B_y = \frac{-\mu I(x+d)}{2\pi((x+d)^2 + y^2)} - \frac{\mu I(x-d)}{2\pi((x-d)^2 + y^2)} \Big|_{y=0} \quad (5.78)$$

Substituting for B_y we have the following easy integral

$$L = \frac{\mu I}{2\pi} \int_a^{d-a} \left(\frac{1}{x} - \frac{1}{x-d} \right) dx \quad (5.79)$$

so that the external inductance per unit length is given by

$$L = \frac{\mu I}{\pi} \ln \left(\frac{d-a}{a} \right) \quad (5.80)$$

The above expression grows logarithmically as a function of separation d . To compute the internal inductance, let us first consider an isolated conductor, or equivalently take $d \gg a$. Then the volume integral of $|\mathbf{B}|^2$ inside the wire is given by

$$\int_V B^2 dV = \int_0^{2\pi} \int_0^a B(r)^2 r dr d\theta \quad (5.81)$$

where

$$B(r)^2 = \frac{\mu^2 I^2 r^2}{4\pi^2 a^4} \quad (5.82)$$

Performing the trivial integral (there aren't too many in this business!), we have

$$\int_V B^2 dV = \frac{\mu^2 I^2}{8\pi} \quad (5.83)$$

and from the energy definition of the inductance, the internal inductance per unit length for the conductor is given by

$$L = \frac{\mu}{8\pi} \quad (5.84)$$

This expression is interestingly independent of the radius of the wire. So for two widely separated wires we double the above result for the internal inductance. The exact solution is more complicated but follows in a similar fashion.

Magnetic vector potential

As we have seen, the concept of the magnetic vector potential comes about directly as a consequence of Maxwell's equations. The magnetic vector **B** is divergenceless

$$\nabla \cdot \mathbf{B} \equiv 0 \quad (5.85)$$

due to the absence of magnetic monopoles. Since $\nabla \cdot \nabla \times \mathbf{F} \equiv 0$ is an identity for any vector function **F**, it follows that

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (5.86)$$

We call the vector **A** the vector potential in analogy with the electric scalar potential ψ . At first, though, it seems like the vector **A** is superfluous since it too is a vector and not a scalar as in the case of ψ . Furthermore there is some physical ambiguity in **A** since its divergence is arbitrary.

The utility in defining **A**, though, is clear when we compare the defining equation for **A** compared with **B**. Under magnetostatic conditions the vector function **A**(**r**) is given by

$$\mathbf{A}(\mathbf{r}) = \mu \int_{V'} \frac{\mathbf{J}(\mathbf{r}')}{4\pi R} dV' \quad (5.87)$$

As before R is the distance between the source and field point and the volume of integration is over all space where $\mathbf{J} \neq 0$.

Compare this equation with the Biot-Savart Law shown in Eq. (5.65). The vector **A** is in the direction of the current and no cross products are involved in the computation. In fact, if we partition the vector $\mathbf{J} = J_x \hat{\mathbf{x}} + J_y \hat{\mathbf{y}} + J_z \hat{\mathbf{z}}$ into Cartesian coordinates, then Eq. (5.87) reduces to three scalar integrals identical in form to the calculation of the electric scalar potential in terms of charge density.

The vector potential **A** is related to inductance as follows. Substitute $\mathbf{B} = \nabla \times \mathbf{A}$ into Maxwell's equation

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \frac{\partial \mathbf{A}}{\partial t} \quad (5.88)$$

Since the curl operator is associative, we have

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = 0 \quad (5.89)$$

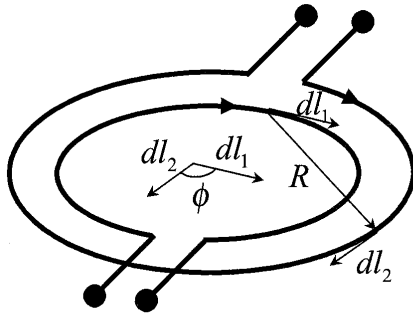


Figure 5.18 Two magnetically coupled filamentary loops.

and this implies that the curl-free expression in the parentheses, similar to the electrostatic field, can be written as the gradient of a scalar function

$$\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla \phi \quad (5.90)$$

The sign of the scalar potential is arbitrarily chosen to correspond to the electrostatic case. In other words, the electric field at any point in space has two components

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \nabla \phi \quad (5.91)$$

If we integrate the above expression around a closed loop, the second term is zero and we are only left with the integral of the first term

$$V_L = \oint_C \mathbf{E} \cdot d\mathbf{l} = -\frac{d}{dt} \oint_C \mathbf{A} \cdot d\mathbf{l} \quad (5.92)$$

leading to an expression for the inductance in terms of the magnetic vector potential

$$L = \frac{1}{I} \oint_C \mathbf{A} \cdot d\mathbf{l} \quad (5.93)$$

We could have derived the above result directly from Stoke's Theorem

$$\int_S \mathbf{B} \cdot d\mathbf{S} = \oint_C (\nabla \times \mathbf{B}) \cdot d\mathbf{l} \quad (5.94)$$

but the above derivation gives us more physical insight into the magnetic vector potential. In fact, Eq. (5.91) justifies the name *potential*. Under general circumstances, the electric field is the gradient of the scalar potential and the time derivative of the vector potential. Both operations involve differentiation. The time derivative with the vector potential reminds us that magnetic voltages are only induced under dynamic conditions. From Eq. (5.91) the SI units of $\mathbf{A}$ must be volt seconds per meter ($\text{V} \cdot \text{s}/\text{m}$, or since the units of $\mathbf{B}$ can be specified in Tesla (T) or weber per square meter (W/m^2), we also have Tesla meter ($\text{T} \cdot \text{m}$) or weber per meter (W/m).

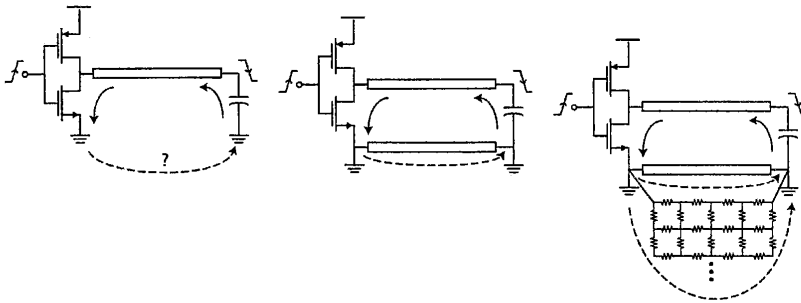


Figure 5.19 (a) An integrated circuit wire carries AC current to a capacitor load. (b) The return current can flow through an explicit path, or (c) through a distributed path formed by the substrate and ground network.

5.5 Partial inductance and return currents

Often people speak of the inductance of an “open” path, such as a piece of wire or a metal trace in an integrated circuit. Quite often, you hear about the inductance of a bond wire or a lead on a package. What are they talking about?

In fact, people are speaking of *partial* inductance. In other words, consider the inductance of the integrated circuit trace shown in Fig. 5.19a. Imagine that we drive this wire with a voltage source at one end through a capacitor load at the other end. Where does the “return current” come from? This depends on how the circuit is laid out. For instance, a separate trace in parallel to the first can carry the return current back to the source, and thus the inductance is defined by this loop (Fig. 5.19b). But often an explicit ground return is not included, and the return current flows through the ground network. Since the substrate is often grounded, the return current can also flow partially through the substrate and thus the inductance is now different since the “loop” is different (Fig. 5.19c). In fact, with the ground and substrate we have a dilemma since we don’t know the exact current distribution. At high frequency, we imagine that most of the current will flow close to the surface in the vicinity of the trace. But at low frequencies, the current flow is more widespread.

Nevertheless, people still say that this trace has “so and so” inductance. They are talking about the inductance of the loop assuming the return current is at infinity. In other words, imagine calculating the inductance through the vector potential as follows. As suggested by Fig. 5.20, if we integrate Eq. (5.93) *partially* through several open paths, then the voltage induced at the terminals of the i th curve is given by

$$V_i = \int_{C_i} \mathbf{E} \cdot d\mathbf{l} \quad (5.95)$$

If we ignore the contribution from ϕ in Eq. (5.91) and only consider the contribution from the vector potential, we then have

$$\hat{V}_i = -\frac{d}{dt} \int_{C_i} \mathbf{A} \cdot d\mathbf{l} \quad (5.96)$$

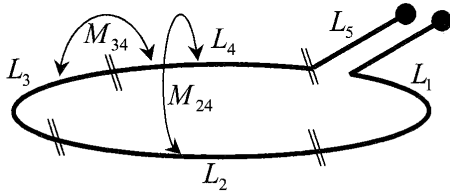


Figure 5.20 Breaking the inductance into partial self and mutual inductance terms.

Due to linearity, we can partition $\mathbf{A}$ into several terms

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 + \cdots \quad (5.97)$$

where each term represents the contribution of segment i to the magnetic vector potential

$$\mathbf{A}_i = \frac{\mu}{4\pi} \int_{V_i} \frac{\mathbf{J}_i}{R} dV_i \quad (5.98)$$

The total current I_i flowing into segment i can be taken out of the integral as long as the current distribution is retained in the integral, $\mathbf{J}(\mathbf{r})_i = \mathbf{k}(\mathbf{r})I_i$. Let us define the *partial* mutual inductance between segment i and j as follows

$$M_{ij} = \frac{1}{I_j} \int_{C_i} \mathbf{A}_j \cdot d\mathbf{l} \quad (5.99)$$

so that the voltage around the entire structure is simply given by

$$V = \sum_i \sum_j M_{ij} \frac{dI_i}{dt} \quad (5.100)$$

From the above definition we define the *partial inductance matrix* as the $N \times N$ matrix M with elements M_{ij} . Given vectors v and i representing the segment voltages and currents, we extend Eq. (5.1)

$$v = M \frac{di}{dt} \quad (5.101)$$

5.6 Impedance and quality factor

The circuit concept of impedance can be stated in terms of power

$$Z_{in} = \frac{V}{I} = \frac{V \cdot I^*}{|I|^2} = \frac{P}{\frac{1}{2}|I|^2} = R + jX \quad (5.102)$$

Applying Poynting's Theorem to a "black box," we can write this as

$$Z_{in} = \frac{P_0 + P_\ell + 2j\omega(W_m - W_e)}{\frac{1}{2}|I|^2} \quad (5.103)$$

Note that the resistive component has a radiation term, an ohmic loss term, and possibly a dielectric or permeability loss term

$$R = \frac{P_0 + P_\ell}{\frac{1}{2}|I|^2} \quad (5.104)$$

The reactance is positive if $W_m > W_e$, and negative otherwise

$$X = \frac{2\omega(W_m - W_e)}{\frac{1}{2}|I|^2} \quad (5.105)$$

The quality factor for a “black box” (usually resonator) is defined as follows

$$Q = 2\pi \frac{\text{Peak energy stored}}{\text{Energy loss per cycle}} \quad (5.106)$$

The denominator can be reformulated in terms of the average power loss to give

$$Q = \omega \frac{\text{Peak energy stored}}{P_\ell} \quad (5.107)$$

From Poynting's Theorem, the net stored energy is given by $W_m - W_e$, so we may be tempted to write

$$Q \stackrel{?}{=} \omega \frac{W_m - W_e}{P_\ell} = \frac{1}{2} \frac{\Im(Z_{in})}{\Re(Z_{in})} \quad (5.108)$$

But the peak energy is different from the *net* energy. In a LC resonator, the energy is shared between the inductor and capacitor in a time-varying fashion. The peak energy can be computed from the capacitor when $i = 0$ or from the inductor when $v_C = 0$.

$$Q = \omega \frac{|W_m| + |W_e|}{P_\ell} = \omega \frac{W_m^{peak}}{P_\ell} = \omega \frac{W_e^{peak}}{P_\ell} = \frac{\omega L}{R} = \frac{1}{\omega C R} = \frac{X}{R} \quad (5.109)$$

For a single one-port element *not* in resonance, we often define the Q factor as

$$Q = \frac{\Im(Z_{in})}{\Re(Z_{in})} = \frac{X}{R} \quad (5.110)$$

We see that this is correct under the assumption that the one-port forms a resonant circuit!

5.7 Frequency response of inductors

Skin effect

At low frequencies when the DC resistance of an inductor dominates over the reactance, for example when $R \gg \omega L$, then current flows evenly through the cross section of a uniform material of a given conductivity. Thus when computing the low-frequency inductance (strictly the DC inductance¹) we have non-zero flux in the volume of conductors and to

¹ The concept of DC may seem meaningless since at DC there is no induction and thus inductance plays no part. However, we shall see that inductance is a weak function of frequency so that even at moderate frequencies, the DC inductance remains valid.

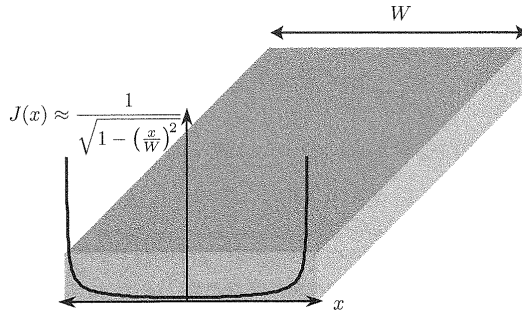


Figure 5.21 Non-uniform current flow through conductors at high frequency. The shown functional dependence is an approximation to the edge discontinuity for rectangular conductors.

calculate the inductance correctly we have to take this into account. At increasing frequency, the internal current distribution redistributes, changing the *internal impedance* of conductors. Ignoring displacement current, the internal impedance is given by

$$Z_{int} = \frac{V_{res}}{I} = \frac{1}{I} \int_C \frac{\mathbf{J}}{\sigma} \cdot d\mathbf{l} \quad (5.111)$$

The above equation is simply the line integral of the electric field $\mathbf{E}$ along the path of a resistive material of conductivity σ . At low frequency when $J = I/A$ is uniform we have the familiar Ohm's Law expression

$$Z_{int} = \frac{1}{I} \int_C \frac{I dl}{\sigma A} = R \quad (5.112)$$

At high frequency $\mathbf{J} = \sigma \mathbf{E}_{tot}$ where $\mathbf{E}_{tot}$ is the total vectorial sum field, including the applied and induced fields. As we alluded to earlier when discussing the internal and external inductance, the induced field is stronger along the outer edge of a loop (which encloses more flux) than the inner loop. Since the induced field opposes the applied field, as shown in Fig. 5.21, we expect the current density to peak near the inner edge of the conductor. In fact, this is exactly what is observed experimentally or by solving Maxwell's equations directly.

We shall show that at increasingly high frequencies the non-uniformity of the current distribution dominates completely and the internal flux is driven to zero. Thus, the infinite frequency asymptote for inductance should equal the value obtained by the external flux

$$L_{\infty} = \frac{\psi_{ext}}{I} \quad (5.113)$$

Note that the DC inductance

$$L_{DC} = \frac{\psi_{ext} + \psi_{int}}{I} \quad (5.114)$$

is larger only by ψ_{int} . Since for usual geometries $\psi_{ext} \gg \psi_{int}$ and $\psi_{int}(f)$ is a decreasing function of frequency, the inductance decreases only slightly with frequency. For instance,

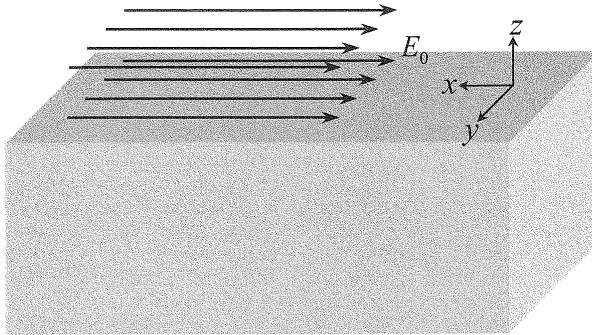


Figure 5.22 Current flow along the surface of a semi-infinite plane.

if the internal flux only accounts for 5% of the total flux, then the high-frequency inductance will only differ by 5% from the DC value.

Surface impedance

In this section we will justify some of the above arguments by solving Maxwell's equations directly. For a simple analytic solution we have to limit ourselves to an unrealistic one-dimensional geometry shown in Fig. 5.22 but the insight afforded by this simple example applies in a surprising number of more general situations.

For a semi-infinite plane supporting a conductive current, we have the following time-harmonic Maxwell equations

$$\nabla \times \mathbf{H} = (\sigma + j\omega\epsilon)\mathbf{E} \quad (5.115)$$

The displacement current $j\omega\epsilon\mathbf{E}$ is negligible in comparison to the conductive current $\sigma\mathbf{E}$. This is true for good conductors and in fact a useful definition for a good conductor. Taking the curl of Maxwell's curl equation for an electric field we have

$$\nabla \times \nabla \times \mathbf{E} = -j\omega\mu\nabla \times \mathbf{H} \quad (5.116)$$

and substituting for the curl of $\mathbf{H}$ from above while dropping the displacement current

$$\nabla \times \nabla \times \mathbf{E} = -j\omega\mu\sigma\mathbf{E} \quad (5.117)$$

The above differential equation describes the spacial variation of the components of the vector $\mathbf{E}$. For simplicity we assume that $\mathbf{E}$ is x -directed so that $\mathbf{E} = \hat{x}E_x$ and furthermore we assume that $\mathbf{E}$ is uniformly incident on the plane. Simplifying the above equation we are led to the following scalar differential equation

$$-\left(\frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2}\right) = -j\omega\mu\sigma E_x \quad (5.118)$$

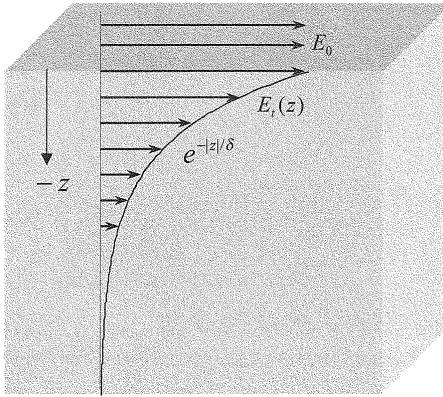


Figure 5.23 The decay of electric field into the depth of a semi-infinite plane conductor.

We assume a uniform wave incident on the plane and thus $\partial E_x / \partial y = 0$ for $z > 0$. By the symmetry of the plane, we can extend this condition to the interior of the plane and we are thus faced with the following simple first-order equation

$$\frac{d^2 E_x}{dz^2} = j\omega\mu\sigma E_x \quad (5.119)$$

The above equation is readily solved by substituting the trial solution $E_0 \exp(\alpha z)$

$$\alpha^2 E_x = j\omega\mu\sigma E_x \quad (5.120)$$

And thus

$$\alpha = \pm \sqrt{j\omega\mu\sigma} = \pm \sqrt{j} \sqrt{\omega\mu\sigma} \quad (5.121)$$

Since the fields must be finite as $z \rightarrow -\infty$, we keep the positive root. Working with polar coordinates, verify that the first term simplifies to $\sqrt{j} = (1 + j)/\sqrt{2}$. Let us define the constant

$$\delta \equiv \sqrt{\frac{2}{\omega\mu\sigma}} \quad (5.122)$$

With this definition we have the following form for the electric field inside the material at depth $z < 0$

$$E_x(z) = E_0 e^{z/\delta} e^{jz/\delta} \quad (5.123)$$

We call δ the *penetration depth* parameter or more commonly the *skin depth*. The motivation for this is clear since δ has units of length (m) and represents the decay rate of the field into the conductor. This is illustrated in Fig. 5.23. Since $\mathbf{J} = \sigma \mathbf{E}$ it is easy to see that the current density decays in a similar manner to the electric field. It can be shown that the magnetic field also diminishes exponentially.

Table 5.1 *The conductivity and penetration depth of various conductors at room temperature*

Material	Conductivity (S/m)	Skin Depth (μ)		
		100 MHz	1 GHz	10 GHz
Copper	5.80×10^7	6.6	2.1	0.66
Aluminum	3.72×10^7	8.2	2.6	0.82
Gold	4.44×10^7	7.6	2.4	0.76
Silver	6.17×10^7	6.4	2.0	0.64
Brass	1.57×10^7	12.7	4.0	1.27

For static fields it is easy to show that $E \equiv 0$ inside perfect conductors. Now we see more generally that AC fields decay to zero in *good* conductors. The rate of decay is quite rapid since σ is large for good conductors. In Table 5.1 we quantify δ for typical materials and some typical frequencies. For instance, for Cu at 1 GHz, the skin depth is only 2.1μ . For this reason, δ is also called the *skin depth* since current flows along the outer skin of good conductors.

It is interesting to note that the phase term in Eq. (5.123) causes the electric field to flip direction inside the material at depth $z = \pi\delta$, thus causing the current to travel “backwards”! What causes this funny behavior? It is the magnetically induced currents. To observe this, note that time-varying conduction current creates a time-varying magnetic field which in turn induces a time-varying current opposing the original current. In the steady state, therefore, the current distribution takes on the form we have described.

It is reasonable to assume that the above solution can be applied to the *interior* of a *finite* rectangular conductor of width w and thickness t as long as the thickness is much greater than the skin depth. We do not expect the solution to apply to the edges of the rectangular conductor and in fact we shall later see that the field behavior at the edges is complicated. A round conductor of large radius (compared with the skin depth) should also have approximately the same field distribution, and, due to the absence of the edge discontinuity, the results should be even more applicable. In the next section we shall have occasion to make this comparison quantitative.

Since the current distribution is non-uniform, we employ Eq. (5.111) to obtain the internal impedance of the rectangular conductor. For the path of integration, we take the path along the surface at $z = 0$, where $\mathbf{E} = E_x \hat{x}$

$$Z_{int} = \frac{1}{I} \int_0^L E_x dx = \frac{E_x L}{I} \quad (5.124)$$

so the internal impedance per unit length, or the *surface impedance* is given by

$$Z_s = \frac{W}{L} Z_{int} = W \frac{E_x}{I} \quad (5.125)$$

The total current is given by integrating $\mathbf{J}$ along the depth of the conductor

$$I = W \int_{-\infty}^0 J(z) dz = W \sigma E_x \int_0^{\infty} e^{-(1+j)z/\delta} dz = W \delta \sigma E_x / (1+j) \quad (5.126)$$

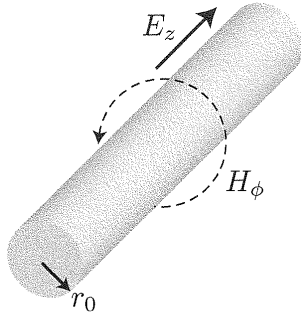


Figure 5.24 A round wire of infinite length has cylindrical symmetry.

Substituting in Eq. (5.125) we have

$$Z_s = \frac{(1+j)}{\sigma\delta} = (1+j)R_s \quad (5.127)$$

Does the above equation look familiar? The real part looks like the *sheet resistance* ($\Omega/\square$) of a material of conductivity σ and thickness δ . In other words, at frequency ω our rectangular slab appears the same as a slab at zero frequency with uniform current flowing through a thickness δ . Thus the resistance of the slab increases $\sqrt{\omega}$ as a function of frequency since $\delta \propto \sqrt{\omega}$. We can also observe from the imaginary part of Z_s that the *internal inductance* decreases at high frequency since the magnetic strength is reduced in the volume of the conductor.

Impedance of round wires

For a long round wire, the current J_z is invariant with z and the angle θ as shown in Fig. 5.24. Therefore, the Helmholtz equation simplifies

$$\nabla^2 \mathbf{J} = j\omega\mu\sigma\mathbf{J} = \tau^2 \mathbf{J} \quad (5.128)$$

$$\frac{\partial^2 J_z}{\partial r^2} + \frac{1}{r} \frac{\partial J_z}{\partial r} + \tau^2 J_z = 0 \quad (5.129)$$

Two linearity independent solutions are the Bessel function and the Hankel function of the first kind

$$J_z = AJ_0(\tau r) + BH_0^{(1)}(\tau r) \quad (5.130)$$

Since the Hankel function has a singularity at $r = 0$, it cannot be a solution. Normalizing to the current at the surface of the wire

$$J_z = \frac{\sigma E_0}{J_0(\tau r_0)} J_0(\tau r) \quad (5.131)$$

A plot of the current density in the wire is shown in Fig. 5.25. In the plot, a Cu wire with 1 mm diameter is used. Note that at low frequencies the current is essentially

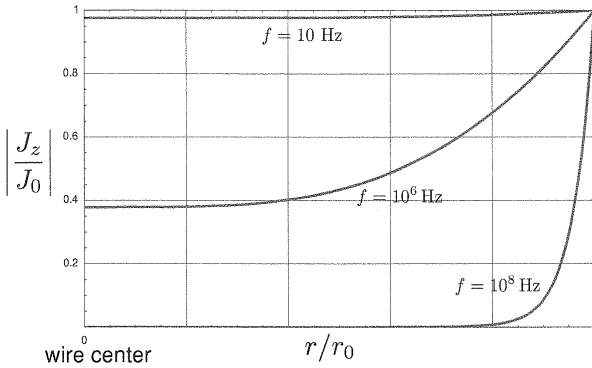


Figure 5.25 The current density inside of a round wire as a function of the radius. The origin represents the center of the wire. The current is uniform at low frequency but concentrated at the skin at high frequency.

uniform. At high frequency, though, the current decays exponentially as we penetrate the conductor.

In the limit that the radius is large, or equivalently $r_0/\delta \gg 1$, then the wire should behave like our plane conductor. In fact

$$\left| \frac{J_z}{\sigma E_0} \right| \approx e^{-(r_0-r)/\delta} \quad (5.132)$$

The impedance of a round wire can be computed by noting that only E_z and H_ϕ are present. Furthermore, we have

$$\oint \mathbf{H} \cdot d\mathbf{l} = I = 2\pi r_0 H_\phi \quad (5.133)$$

By $\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}$, it is easy to show that

$$H_\phi = \frac{1}{j\omega\mu} \frac{\partial E_z}{\partial r} \quad (5.134)$$

Using $E_z = J_z/\sigma = E_0 J_0(\tau r)/J_0(\tau r_0)$, the magnetic field is given by

$$H_\phi = \frac{E_0 \tau}{j\omega\mu} \frac{J'_0(\tau r)}{J_0(\tau r_0)} \quad (5.135)$$

Recall that $J'_0(x) = -J_1(x)$. Solving for the current

$$I = \frac{2\pi r_0 \sigma E_0}{\tau} \frac{J_1(\tau r_0)}{J_0(\tau r_0)} \quad (5.136)$$

Finally, we can write the internal impedance of the wire

$$Z_i = \frac{E_z(r_0)}{I} = \frac{\tau J_0(\tau r_0)}{2\pi r_0 \sigma J_1(\tau r_0)} \quad (5.137)$$

The above expression involves the Bessel function and its derivative of complex argument. While these functions have been amply tabulated, any modern computation environment such as *matlab* or *mathematica* can be used to evaluate the above expression.

In the low-frequency limit, the internal impedance of the wire reduces to

$$Z_i \approx \frac{1}{\pi r_0^2 \sigma} \left[1 + \frac{1}{48} \left(\frac{r_0}{\delta} \right)^2 \right] + j\omega \frac{\mu}{8\pi} \quad (5.138)$$

The real part corresponds to a correction to the DC resistance of the wire (per unit length). The imaginary term corresponds exactly to the static internal inductance of the wire. As expected, the high-frequency limit matches the analysis for the semi-infinite plane

$$Z_i = \frac{(1+j)R_s}{2\pi r_0} \quad (5.139)$$

This is because at high frequency the perimeter of the conductor is much larger than the penetration depth δ and thus the plane conductor analysis becomes valid.

Approximate impedance of rectangular wires

The resistance of a rectangular wire as a function of frequency can be obtained numerically and fitted to the following equation [25]

$$R = \frac{l}{\sigma w t} \left[\frac{0.43093 x_w}{1 + 0.041 \left(\frac{w}{t} \right)^{1.19}} + \frac{1.1147 + 1.2868 x_w}{1.2296 + 1.287 x_w^3} + 0.0035 \left(\frac{w}{t} - 1 \right)^{1.8} \right] \quad (5.140)$$

where $x_w = \sqrt{2f\sigma\mu w t} \geq 2.5$ and for $x_w < 2.5$

$$R = \frac{l}{\sigma w t} \left[1 + 0.0122 x_w^{3+0.01 x_w^2} \right] \quad (5.141)$$

The above expression has an accuracy between 3% and 5% in the range of $w/t < 12$ and $x_w < 20$. It is interesting to compare the above result with the surface impedance calculation to see how much the edge effects alter the resistance.

Lossless inductors

The above discussion only applies to materials with finite conductivity. For lossless materials, the fields diminish in the volume of conductors and an infinitely dense current flows on the surface of conductors. The tangential electric field is zero inside and just outside of the conductors. The tangential magnetic field H_t is also zero inside of the volume of conductors but has a jump discontinuity equal to the surface current density J_s just outside the space of conductors

$$H_t = J_s \quad (5.142)$$

Thus the flux is independent of frequency and we do not expect the same variation in inductance due to the skin effect. On the other hand, the effective inductance will vary due to capacitive effects.

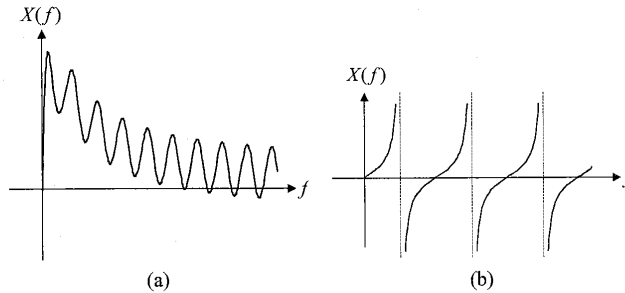


Figure 5.26 (a) The reactance of a lossy passive device as a function of frequency. (b) The reactance of a lossless passive device as a function of frequency.

Consider the complex power flowing into a black box shown in Fig. 5.9. If we integrate the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}^*$ across a surface S enclosing the black box

$$P = \frac{1}{2} \oint_S \mathbf{E} \times \mathbf{H}^* \cdot d\mathbf{S} = P_l + 2j\omega(W_m - W_e) \quad (5.143)$$

From electromagnetic theory [52] we have written the imaginary part of the above expression as the difference between the magnetic and electric stored energy. The reactance of the black box is thus related to the energy stored as follows

$$X = \frac{4\omega(W_m - W_e)}{II^*} \quad (5.144)$$

For an inductor $W_m > W_e$ and ideally $W_m \gg W_e$ so that the reactance of an inductor is positive. Furthermore it can be shown [52] that

$$\frac{dX}{d\omega} = \frac{4(W_m + W_e)}{II^*} \quad (5.145)$$

Please note that the above result is not a simple derivative of Eq. (5.144) (notice the change in sign). This shows that the derivative of the reactance is positive. This is a surprising result since it says that a lossless inductor cannot have a frequency-dependent profile such as shown in Fig. 5.26a, but it must resemble Fig. 5.26b. In other words, the reactance of an inductor is positive and *always* increasing as a function of frequency. The inductance as a function of frequency

$$L(\omega) = \frac{X}{\omega} \quad (5.146)$$

has slope

$$\frac{dL}{d\omega} = \frac{1}{\omega} \frac{dX}{d\omega} - \frac{X}{\omega^2} \quad (5.147)$$

Notice that every term in the above equation is positive for $\omega > 0$. From circuit theory, at low frequencies we expect

$$\frac{dX}{d\omega} = \frac{X}{\omega} \quad (5.148)$$

so that $dL/d\omega = 0$, implying a constant inductance. If the slope of the reactance is ever smaller than the reactance divided by frequency, the inductance will have a negative slope. But comparing Eq. (5.144) to Eq. (5.145), it is clear that $\frac{dX}{d\omega} \geq \frac{X}{\omega}$ so that

$$\frac{dL}{d\omega} \geq 0 \quad (5.149)$$

It is important to qualify the above result. First, we have assumed lossless materials in the construction of the inductor. Second, we have defined inductance by (5.146). Earlier we defined the inductance in terms of the magnetic energy alone. We should really call this the *effective* inductance since at high frequencies it will contain both magnetic and electrical energy. It is nevertheless interesting to notice that the effective inductance has a positive slope as a function of frequency somewhat counterintuitively. This result, though, is consistent with circuit theory. If we calculate the effective inductance of an LC tank circuit, we have

$$L_{eff} = \frac{L}{1 - LC\omega^2} \quad (5.150)$$

and for $\omega < \omega_0 \equiv 1/\sqrt{LC}$ the inductance is positive and increasing. As $\omega \rightarrow \omega_0$, the effective inductance takes on arbitrarily high values.

5.8 Quality factor of inductors

Using the fundamental definition of the Q factor as the ratio of energy storage to energy loss in a practical inductor over a cycle of excitation

$$Q \equiv \frac{\text{energy stored}}{\text{energy loss}} \quad (5.151)$$

For a physical inductor that has resistance R at frequency ω , in one cycle the inductor dissipates

$$E_{loss} = \frac{1}{2}|I|^2 RT = \frac{1}{2}|I|^2 R \frac{2\pi}{\omega} \quad (5.152)$$

Similarly, the energy stored in an inductor is

$$E_{stored} = \Im \left[\frac{1}{2} IV^* T \right] = \frac{1}{2}|I|^2 \omega L \frac{2\pi}{\omega} = \frac{1}{2}|I|^2 L 2\pi \quad (5.153)$$

and thus

$$Q = \frac{\omega L}{R} \quad (5.154)$$

The quality factor of inductors is important in many applications. For instance, a resonant tank exchanges energy between the magnetic and electric fields every cycle. If a fraction $1/Q$ of the stored energy is lost every cycle, then the system can sustain itself for about Q cycles. This determines, for instance, the power that must be replenished to the tank to sustain oscillations. While active devices such as transistors can be used to do this, these devices always add undesirable noise to the system.

To improve an inductor, therefore, we must minimize the various loss mechanisms associated with the inductor.

Inductor loss mechanisms

Conductive losses

At low frequencies, the resistance of an inductor is simply related to the DC resistance of the windings that make up the inductor. Since the resistance is an increasing function of temperature, the on-chip temperature should be estimated in determining the losses.

Internally induced losses

At high frequencies the effective resistance of an inductor is complicated by non-uniform current distribution in the interior of conductors. We have already seen that the *skin effect* tends to decrease the cross-sectional area of the device and thus we expect a resistance increase proportional to $\sqrt{f}$ due to this effect alone.

The physical asymmetry in the geometry of the structure will only exacerbate this. For instance, we observed that in a loop inductor the current tends to crowd to the inner radius of the conductor.

Externally induced losses

The proximity of nearby conductors also influences loss in an inductor. Due to the time-varying flux, *eddy currents* are induced to flow in conductors. As shown in Fig. 5.27, the effective resistance of these eddy current loops appears in series with the inductor. If we assume that the magnetic field of the induced currents interacts weakly with the magnetic field of the inductor, then we can neglect the change in inductance due to these currents and calculate the eddy currents directly as $j\omega\mathbf{A}$, where $\mathbf{A}$ is determined by the inductor currents alone. Otherwise, we must compute $\mathbf{A}$ by including the inductor current as well as the induced currents and this is a complicated situation.

Magnetization losses

Hitherto we have implicitly assumed that the material properties are linear, isotropic, and real. For most non-magnetic or weakly magnetic materials these are good assumptions. For magnetic materials, though, the magnetic permeability μ is non-linear and frequency dependent and dissipative. The exact physical cause of this is beyond the scope of this chapter

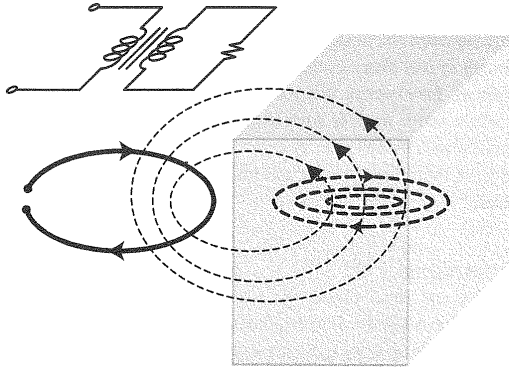


Figure 5.27 Model for eddy current losses as effective series losses.

and requires analysis of the quantum mechanics and thermodynamics of the constituent atoms. We can sweep such details under the rug and assume that $\mu = \mu' + j\mu''$ is a complex frequency-dependent quantity due to losses. The Q factor is thus bounded by μ'/μ'' due to the lossy part of the volume integral of $\mathbf{B} \cdot \mathbf{H}$ over all space.

Displacement current losses

Inevitably every real inductor stores some electric energy. We have up to now neglected the small displacement current $\mathbf{J}_d = j\omega\epsilon\mathbf{E}$ that flows in the volume of the problem space. If the material in question is lossy, then the permittivity has an imaginary component $\epsilon = \epsilon' - j\epsilon''$ as well. Physically, the electrically induced polarization in the material drains energy from the field and thus acts like a loss mechanism. The ratio of the imaginary part of the permittivity to the real part is known as the *loss tangent*

$$\tan \delta = \frac{\epsilon''}{\epsilon'} \quad (5.155)$$

and this is specified for various materials at various frequencies. We will defer a detailed discussion of the loss tangent. For now we will merely point out that the loss tangent acts as an additional loss mechanism when an inductor stores electric energy in addition to magnetic energy.

Radiative losses

For inductors that are large relative to the wavelength $\lambda = c/f$, additional losses can occur. For instance, consider breaking the inductor into many small sections and employing the partial inductance concept, the induced field at segment i due to current in segment j is simply $j\omega\mathbf{A}_j$. When segments i and j are in close proximity, the vector $\mathbf{A}$ is purely real and so the induced field only stores energy. For points far removed, though, there is a

propagation delay between these two points which adds an additional phase to A_j when evaluated at segment i . This additional phase delay causes the induced fields to have a *real* component, which appears as additional resistance. Physically, energy is lost to *radiation*, an electromagnetic wave which travels into space.

Overall quality factor due to multiple loss mechanisms

Consider the basic definition of the Q factor as the ratio of energy storage to energy loss. If we expand the various loss mechanisms we have

$$Q = \frac{W_{\text{stored}}}{W_{\text{loss}}} = \frac{W_{\text{stored}}}{W_{\text{loss},1} + W_{\text{loss},2} + \dots} \quad (5.156)$$

Taking the reciprocal of the above we have

$$\frac{1}{Q} = \frac{W_{\text{loss},1} + W_{\text{loss},2} + \dots}{W_{\text{stored}}} = \frac{1}{Q_1} + \frac{1}{Q_2} + \dots \quad (5.157)$$

And thus the overall Q can be computed as the “parallel” combination of the individual Q s. The overall Q factor is therefore limited by the smallest Q factor and is always less than this quantity. As expected, when one loss mechanism dominates, the Q factor is approximately determined by this loss term. For instance, for an inductor constructed over an insulating substrate (such as GaAs) the displacement current losses may be orders of magnitude smaller than the ohmic losses and can be safely neglected. This is in stark contrast to Si, where the conductive substrate contributes significantly to the loss at high frequency.

5.9 Inductors and switching circuits

Ideal inductors are DC shorts. With the absence of resistive loss, the voltage drop along the path of an inductor is equal to the time rate of change of flux, or

$$V_L(t) = \frac{d\psi}{dt} = L \frac{dI}{dt}$$

which is zero for static fields. More interesting, though, is that this implies that the average voltage across an inductor is almost always zero. To see this, consider the Laplace Transform of the voltage. The average value can be obtained by taking the limit as $s \rightarrow 0$. Assuming the Laplace Transform of the current does not have any zeros we have

$$V_L(s \rightarrow 0) = \lim_{s \rightarrow 0} I_L s L = 0 \quad (5.158)$$

A similar conclusion applies for period excitation if we replace $s = j\omega$. Thus the average voltage across an inductor is zero. Therefore, for most periodic waveforms across an inductor, the average value of the steady-state voltage inductor must be zero.

As an application of this concept, consider Fig. 5.28a where an inductor terminal is periodically shorted to ground by an ideal switch with a duty cycle of say 50%. What is the steady-state voltage at node v_1 as a function of time? For 50% of the period, the voltage v_1 is zero due to the ideal switch applying the full rail voltage across the inductor terminals.

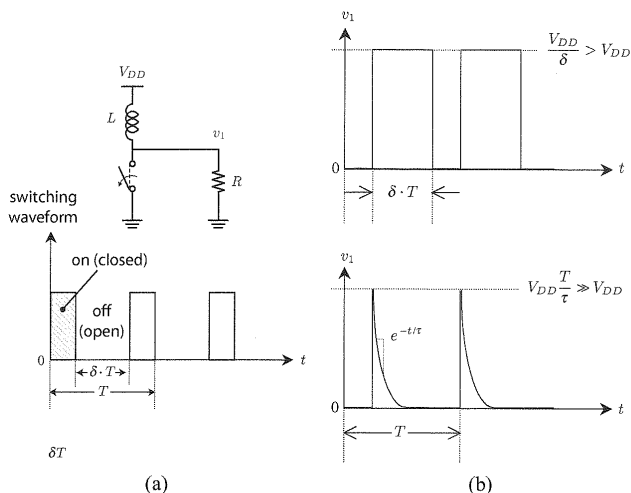


Figure 5.28 (a) Simple LR circuit switching circuit. (b) Approximate voltage waveforms at node v_1 for the case of $\tau \gg 1$ (top) and $\tau \ll 1$ (bottom) where $\tau = L/R$ is the characteristic time constant of the circuit.

In order to maintain an average value of zero, then the voltage v_1 must increase to at least a voltage of $2 \times V_{dd}$ for the remaining off cycle, a voltage above the supply! In fact, if we assume that the inductor is sufficiently large such that $\tau = L/R \gg 1$, then we will observe a steady-state voltage waveform similar to the top plot of Fig. 5.28b. In this case the energy stored in the inductor is sufficiently large that the energy drain during the off-period does not significantly reduce the magnetic energy of the inductor. Stated differently, the inductor is able to maintain a constant current.

On the other hand, if the inductor is small such that $\tau = L/R \ll 1$, then the magnetic energy in the inductor will exponentially drain and the steady-state voltage waveform will resemble the bottom graph of Fig. 5.28b.

Since the output voltage is larger than the supply voltage, this simple circuit principle can be used to build a voltage multiplier. In fact, as the duty cycle of the on period is increased, the voltage during the off-period has to take on correspondingly larger values to sustain a zero average voltage. By using feedback, we can generate a constant voltage by regulating the duty cycle of the waveform to maintain a constant DC voltage over the changing loads.

While at DC, inductors are shorts, at AC they in fact look like “bad” current sources. Just as capacitors at high frequency look like bad voltage sources, inductors have the dual behavior. A constant current drawn from an inductor produces no voltage drop. Inductors present increasingly more reactance to any attempts at changing the current. In many high-frequency circuits this behavior is exploited and large inductors, or RF chokes, are employed to deliver bias currents without disturbing the high-frequency behavior of circuits.

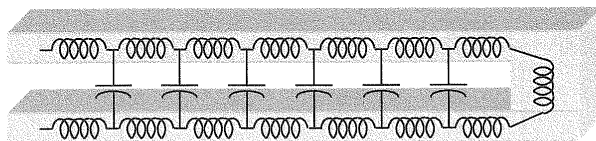


Figure 5.29 The capacitive coupling in a simple loop inductor.

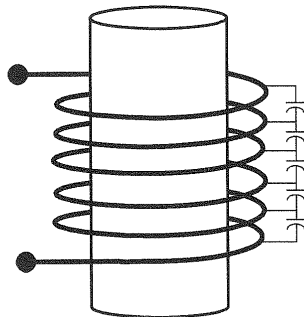
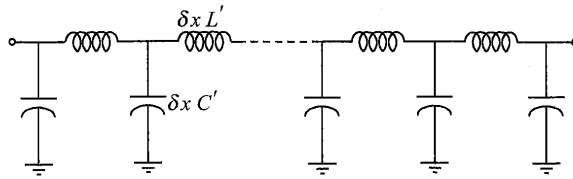


Figure 5.30 The capacitive coupling in a coil inductor.

5.10 Preview: how inductors mutate into capacitors

In Section 5.7 we observed that a structure appears inductive as long as the stored magnetic energy dominates over the electrically stored energy, $W_m > W_e$. Consider the loop inductor shown in Fig. 5.29. Schematically we have shown two paths between the terminals, a *conductive* path and a *displacement* path. The displacement path occurs by the time-varying electric field between the conductors. This electric field only exists if there is a net charge buildup on the two “plates” of the capacitors. It is important to observe that the actual displacement and conductive current paths are distributed and occur side by side. Thus our structure is really both an inductor and a capacitor. In Fig. 5.30 we have shown a common coil inductor and some representative capacitive elements that couple energy electrically from turn to turn of the inductor. While the coupling between adjacent turns is shown, every turn couples to every other turn in the coil.

Define the inductance per unit length as L' and the capacitance per unit length as C' . At low frequencies the loop presents a low impedance since $\omega L'$ is small, whereas $1/(\omega C')$ is large and thus the bulk of the current flows through the loop. Overall the behavior appears as inductive. At high frequencies, though, the inductive reactance is increasing, whereas the capacitive reactance is decreasing, so that there will be a frequency point when $1/(\omega C') = \omega L'$. At this point we may guess that the structure will no longer behave as an inductor and will in fact mutate into a capacitor. While the overall conclusion is correct, the frequency when this occurs is predicted incorrectly due to distributed effects.

Figure 5.31 A distributed LC line.

But at what frequency does the structure *resonate*? At resonance the energy stored in the magnetic field is equal to the energy stored in the electric field. The answer to the question comes from *distributed* analysis of the structure. For instance, the loop inductor of Fig. 5.31 can be analyzed with transmission line theory (Chapter 9). The structure shown can be modeled as a shorted transmission line. Defining the inductance per unit length L' and the capacitance per unit length C' , the impedance of a shorted transmission line is

$$Z_s = j Z_0 \tan(\omega l / c) \quad (5.159)$$

where l is the length of the line, $c = \frac{1}{\sqrt{L'C'}}$ is the propagation velocity, and $Z_0 = \sqrt{\frac{L'}{C'}}$ is the characteristic impedance of the line. Thus the impedance of the line is inductive until $\omega l / c < \pi/2$. But the tangent function is periodic and thus there are an infinite number of frequency ranges where the above device appears inductive. In terms of the inductance and capacitance per unit length, the circuit resonates at frequency ω if

$$n \frac{\pi}{2} = \beta l = \omega \sqrt{L'C'} l = \omega \sqrt{LC} \quad (5.160)$$

for n an integer. Solving for ω

$$\omega = n \frac{\pi}{2} \sqrt{\frac{1}{LC}} \quad (5.161)$$

The above result contradicts circuit theory. Circuit theory predicts a single resonant frequency at a frequency $2/\pi$ lower than distributed analysis. This is not surprising since circuit theory neglects any spacial variation in the fields. We shall return to this subject in Chapter 11. Since any physical inductor stores some parasitic capacitive energy, no matter how small, there is a frequency limit to the inductive behavior.

5.11 References

The ideas for this chapter come from working in the field for many years and so it is hard to trace the material back to the source. I know many of the ideas were inspired by Feynman's Lectures [18], and a careful reading of *Fields and Waves in Communication Electronics* [52]. In fact, the idea of writing this book is partly a result of reading Chapter 4 of this book and wanting to turn the chapter into an entire book.